AD_________________

Award Number: **W81XWH-08-1-0731**

TITLE: **Grid-Enabled Quantitative Analysis of Breast Cancer**

PRINCIPAL INVESTIGATOR: **Andrew R. Jamieson**

CONTRACTING ORGANIZATION:

University of Chicago
Chicago, IL 60637

REPORT DATE: **October 2010**

TYPE OF REPORT: **Annual**

PREPARED FOR: **U.S. Army Medical Research and Materiel Command**
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:

☒ Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
01-10-2010

2. REPORT TYPE
Annual

3. DATES COVERED *(From - To)*
1 Oct 2009 - 30 Sep 2010

4. TITLE AND SUBTITLE
Grid-Enabled Quantitative Analysis of Breast Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-08-1-0731

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Andrew R. Jamieson

Email: andrewj@uchicago.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Chicago

Chicago, IL 60637

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Material Command
Fort Detrick, Maryland 21702-

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

See Next Page.

15. SUBJECT TERMS
Multi-Modality Breast Image Analysis, Grid Computing, Dimension Reduction and Representation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	72	**19b. TELEPHONE NUMBER** *(include area code)*

The long-term goal of this research is to improve breast cancer diagnosis, risk assessment, response assessment, and patient care via the use of large-scale, multi-modality computerized image analysis. The central hypothesis of this research is that large-scale image analysis for breast cancer research will yield improved accuracy and reliability when optimized over multiple features and large multi-modality databases. In the first year of research, we designed a pilot study utilizing large scale parallel Grid computing harnessing nationwide infrastructure for medical image analysis. Also, using a 256-CPU high-throughput computing cluster, dimension reduction techniques were applied to ultrasound, full-field digital mammography, and DCE-MRI breast CADx feature spaces. Results indicated the ability to rival or exceed traditional breast CADx performance. Building on this success, during the second year, we investigated methods for using unlabeled ("truth-unknown") data. Often, there are practical difficulties in assembling large, labeled (histo-pathology) breast image data sets, while unlabeled data may be abundant. This is problematic for conventional CADx schemes reliant on supervised classifiers trained using labeled data only. We proposed using unlabeled breast image data to enhance breast CADx. We hypothesize that unlabeled data information can act as a "regularizing" factor aiding classifier robustness. After conducting experiments using previously collected data sets, encouraging results were found indicating unlabeled data can improve CADx classifier performance.

Table of Contents

Page

INTRODUCTION

Breast cancer is a leading cause of death in women, causing an estimated 40,000 deaths per year.[1] Mammography is the most effective method for the early detection of breast cancer, and it has been shown that periodic screening of asymptomatic women does reduce mortality.[2] Many breast cancers are detected and referred for surgical biopsy on the basis of a radiographically detected mass lesion or cluster of microcalcifications. Useful interpretation in mammography depends on the quality of the mammographic images and the ability of the radiologists who interpret those images.[3] In addition to mammography, follow up imaging with the use of other modalities, such as MRI, and ultrasound are used for assessing malignancy of objects discovered following routine screenings. The long-term goal of this research is to improve breast cancer diagnosis, risk assessment, response assessment, and patient care via the use of large-scale, multi-modality computerized image analysis. The central hypothesis of this research is that large-scale image analysis for breast cancer research will yield improved accuracy and reliability when optimized over multiple features and large multi-modality databases. In recent years, data mining and data driven discovery have become important research tools in many disciplines. Massive amounts of data may contain hidden structure and rich information, previously unavailable for characterization within smaller subgroups. Systematic search can reveal that structure and information. In our context, the opportunity is as follows: The digital age of medical imaging provides an ever-growing archive of data. Deep analysis of this multimodal imaging data can be used to train and optimize algorithms that are incorporated into usable clinical systems, thus improving overall breast imaging interpretation and patient outcome. Data mining can also enable relational discoveries between image data and cancer diagnosis, response, and outcome, thus adding to the potential for "patient-specific diagnoses leading to patient-specific management." Aspects of optimization in this process of CADx development, were previously infeasible due to massive data and computation requirements. However, now with advances in Grid-based computing many research avenues exist.

This second annual report covers the continued developments in research accomplished in the past year leading towards these long term objectives.

BODY

Research Accomplishments

For continuity and completeness, we first list our previous research accomplishments leading up to current progress. Detailed summaries are provided in the attached materials as indicated below.

Recapitulation of Research for Year 1: Oct. 2008 - Oct. 2009

1. Designed and Successfully Executed Proof of a Principle Grid-based Breast CADx Images Analysis Work-flow using Swift Script.

We designed and executed a pilot study to utilize large scale parallel grid computing to harness the nationwide cluster infrastructure for optimization of medical image analysis parameters. **.** Using the grid-environment workflow, parameter sweeps were conducted for lesion segmentation settings based on radial-gradient-index (RGI) methods. Specifically, the Gaussian width (GW) used in initially filtering lesion images for segmentation was varied by increments of 1 mm from 1 to 60 mm. For each GW sweep the entire 850 biopsy-proven mass lesion database (411 benign, 439 malignant) was analyzed. In each, 29 different mathematical descriptor features were calculated, followed by feature selection and merging with linear discriminate analysis. Diagnostic performance was estimated by ROC analysis by calculating AUC values based on both individual features alone, and merged. For merged classifiers, AUC values were found using round-robin case-by-case removal and replacement. Among the resulting , computation jobs requiring over 30 CPU hours on a single lab computer were completed in approximately 35 minutes in this preliminary study. Merged AUC values increased from 0.50 (std.err.=0.018) at GW of 1mm with, to 0.81 (std.err.=0.015) at 10mm GW, with relative plateaus across the rest of the parameter space to 60mm. See conference poster reproduced in **Appendix A** for more details and attached conference poster. [A.R. Jamieson, M L Giger, M Wilde, L Pesce, I Foster, “Grid-Computing for Optimization of CAD,” *Poster,* 50th Assembly and Annual Meeting of American Association of Physicist in Medicine, Houston, Illinois, USA, July 2008]

2. Investigation and Evaluation of Dimension Reduction(DR) in Place of Feature Selection Breast CADx

We applied recently-developed unsupervised non-linear dimension reduction (DR) and data representation techniques to computer-extracted breast lesion feature spaces across three separate imaging modalities: ultrasound (US) with 1126 cases, dynamic contrast enhanced-magnetic resonance imaging (DCE-MRI) with 356 cases, and full-field digital mammography (FFDM) with 245 cases. For the high-dimensional feature spaces, DR methods were tested across all modalities for a range of lower target dimensions and user-defined algorithm parameters. We evaluated the classifier performance using the area under the Receiver Operating Curve ROC curve (AUC) and statistical re-sampling validation techniques. The new techniques were compared to previously developed breast CADx methodologies, including Automatic Relevance Determination (ARD) and linear step-wise (LSW) feature selection, as well as a linear DR method based on Principal Component Analysis (PCA). Using ROC analysis and 0.632+ bootstrap validation, 95% empirical confidence intervals were computed for the each classifier's AUC performance. In the large US dataset, sample high performance results include, $AUC_{0.632+}$ = 0.88 with 95% empirical bootstrap interval [0.787;0.895] for 13 ARD selected features and $AUC_{0.632+}$ = 0.87 with interval [0.817;0.906] for 4 LSW selected features compared to 4D t-SNE mapping (from the original 81D feature space) giving $AUC_{0.632+}$ = 0.90 with interval [0.847;0.919], all using the MCMC-BANN. In conclusion, our preliminary results appear to indicate capability for the new methods to match or exceed classification performance of current advanced breast lesion CADx algorithms. While not appropriate as a complete replacement of feature selection in CADx problems, DR techniques offer a complementary approach which can aid elucidation of additional properties associated with the data. Please see **Appendix B, Results - section IV.A, Figure 1** for complete summary of results, as published in the peer-reviewed *Medical Physics* journal,.[4] [A.R. Jamieson, M. L. Giger, et. al. "Exploring Non-Linear Feature Space Dimension Reduction and Data Representation in Breast CADx with Laplacian Eigenmaps and t-SNE," Medical Physics. 37, 339 (2009).]

3. Investigation Breast CADx Feature Data Representation and Visualization with Non-Linear Local Geometry Preserving Dimension Reduction Methods

We used the output from the unsupervised non-linear dimension reduction (DR) and data representation techniques to visually perceive the feature space data structure across modalities including ultrasound (US) with 1126 cases, dynamic contrast enhanced-magnetic resonance imaging (DCE-MRI) with 356 cases, and full-field digital mammography (FFDM) with 245 cases. Specifically, the new techniques were shown to possess the added benefit of delivering sparse lower-dimensional representations for visual interpretation, revealing intricate data structure of the feature space**.** These visual results for low dimensional visualization of initially high-dimensional CADx feature are provided in **Appendix B, Results - section IV.B, Figure 3 and Figure 4**. [4] [A.R. Jamieson, M. L. Giger, et. al. "Exploring Non-Linear Feature Space Dimension Reduction and Data Representation in Breast CADx with Laplacian Eigenmaps and t-SNE," Medical Physics. 37, 339 (2009).]

4. Initial Investigation of Manifold Regularization for Breast CADx using Unlabeled Image Data

Previously, we began preliminary consideration of techniques for incorporating unlabeled data. As described below, during the second year of research, these ideas were more fully developed and experimentally evaluated.

New Research for Year 2: Oct. 2009 - Oct. 2010

1. Design Breast CADx Scheme that Use Unlabeled Data

Learning with unlabeled data relies on two main assumptions.[5] First, we assume unlabeled data is drawn from the same underlying population as the labeled data used for classifier training. Second, knowledge limited to the marginal probability distribution, P_x (i.e. without labeling), contributes to identifying the class conditional probability, $P(y|x)$ where y is the target class label. Essentially, this requires that if two points, x_1 and x_2 are close according to the intrinsic geometry of P_x, the conditional probabilities $P(y|x_1)$ and $P(y|x_2)$ are likely to be similar. In general, learning with both labeled and unlabeled data is called semi-supervised learning. The concepts are described below and further illustrated in **Appendix C, Introduction -- section I, and section I, Figure 1.** [A.R. Jamieson, ML Giger, et.al. "Enhancement of breast CADx with unlabeled data," Medical Physics 37, 4155 (2010).]

Algorithms for using unlabeled data

Feature extraction is identical for labeled and unlabeled cases. Thus, information from unlabeled and labeled cases can be combined using unsupervised dimension reduction. Ideally, the unlabeled data can help to more accurately capture the underlying manifold structure of the population of imaged objects. We hypothesized that a supervised classifier trained on the labeled data sub-space produced by this type of reduced mapping (i.e., when including unlabeled data during dimension reduction) could lead to enhanced classification performance. We call this approach transductive dimension reduction regularization (TDRR). [6] Unfortunately, to classify new cases, the TDRR approach requires computing new mappings (and consequently classifier re-training). To avoid this problem, an alternative approach for incorporating unlabeled data is Manifold Regularization (MR).[5] Manifold Regularization is considered a “truly” semi-supervised learning (SSL) technique since it can classify new, “out-of-sample” cases without re-training. Manifold Regularization works by minimizing a regularized loss function which includes a term for penalizing decision functions that are not smooth relative to the intrinsic geometry of the data structure (including unlabeled points). Similar to Laplacian Eigenmaps (see **Appendix B, section III.D.i**), the intrinsic geometry of the data is estimated using the graph Laplacian. Diagrams illustrating the different CADx schemes can be found in **Appendix C, section II.B, Figure 2, section III.C.ii, Figure 3.** A more detailed explanation of breast CADx schemes for incorporating unlabeled data can be found in **Appendic C, Methods - section III.C.** . [A.R. Jamieson, ML Giger, et.al. "Enhancement of breast CADx with unlabeled data," Medical Physics 37, 4155 (2010).].

2. Experimentally Evaluate Breast CADx Schemes using Unlabeled Data

We investigated the use of three unsupervised dimension reduction techniques (PCA, Laplacian Eigenmaps, and t-SNE) in the first stage of the TDRR scheme, coupled with a Bayesian Neural Net (BANN) supervised classifier in the second stage.[7,8] The dual-stage TDRR scheme was compared to a single-stage scheme based on Manifold Regularization implemented via the LapSVM algorithm.[5] Experiments were conducted using randomly sampled subsets pulled from a relatively large, previously acquired labeled ("truth-known") ultrasound dataset (1126 cases). We hypothesized that the two most important factors influencing performance are the number of cases and the prevalence of cancer (both for the labeled and unlabeled samples). Within constraints imposed by limited data, our experiments attempted to mimic clinically relevant scenarios. Cancer prevalence was fixed at 50% malignant for labeled samples and 5% malignant for unlabeled (other prevalence configurations were investigated but not included here). For the supervised training and testing we focused on smaller set sizes of fifty (50L), one hundred (100L), and one hundred fifty (150L) labeled lesions. Because of high computation demand, we explored only a limited number of unlabeled dataset sizes (three): small, medium, and as large as possible (up to 900UL cases). For each experimental configuration, 200 independently randomly sampled (without replacement) sub-sets were drawn from the entire dataset and identified to the algorithm as labeled or unlabeled accordingly. Labeled and unlabeled subset cases were always mutually exclusive. Further details can be found in **Appendix C, Methods section III.D** [A.R. Jamieson, ML Giger, et.al. "Enhancement of breast CADx with unlabeled data," Medical Physics 37, 4155 (2010).]

Classification performance was estimated by leave-one-out (LOO) cross-validation for the 50L and 100L experiments, 0.632+ bootstrap (632+) for the 150L experiments, and using an independent test set (with 101 lesions), obtained separately from the original dataset. For each of the 200 runs in an experiment, the ΔAUC was computed (ΔAUC defined as AUC with unlabeled - AUC without unlabeled). The paired, non-parametric Wilcoxon signed-rank test was applied to the full set of AUC values and also to quartile sub-groups ordered by the original AUC (without unlabeled), i.e., the 25th percentile, 25th to 50th, 50th to 75th, and the 75th to 100th percentile. P-values were adjusted for multiple comparisons testing using the Holm-Sidak step-down method.[9,10] **Appendix C, Methods section III.E.** [A.R. Jamieson, ML Giger, et.al. "Enhancement of breast CADx with unlabeled data," Medical Physics 37, 4155 (2010).]

Statistically significant differences in the average AUC, between training with and without unlabeled data were detected (i.e., ΔAUC ~= 0). For example, when training using 100 labeled and 900 unlabeled cases and testing on the independent set, the TDRR method using t-SNE produced an average ΔAUC = 0.0361 with 95% intervals [0.0301; 0.0408] and when using Laplacian Eigenmaps an average ΔAUC = .026 [0.0227, 0.0298], while the Manifold Regularization based LapSVM produced an average ΔAUC = .0381 [0.0351; 0.0405] (all with p-values << 0.0001, adjusted for multiple comparisons, but considering the test set fixed). As expected, the linear PCA TDRR scheme did not show improved performance with unlabeled data (detailed results are found in **Appendix C, Results – section IV, Table 4a, 4b**).[6] [A.R. Jamieson, ML Giger, et.al. "Enhancement of breast CADx with unlabeled data," Medical Physics 37, 4155 (2010).] Additionally, we observed that schemes obtaining initially lower than average performance when using labeled data only, showed the most prominent increase in performance when unlabeled data were added, suggesting a regularization effect. (detailed results are found in **Appendix C, Results – section IV, Figure 7, Figure 8**). [6] [A.R. Jamieson, ML Giger, et.al. "Enhancement of breast CADx with unlabeled data," Medical Physics 37, 4155 (2010).] Preliminary findings appear to support our hypothesis that unlabeled data can enhance breast CADx performance by non-negligible amounts. We believe further investigation is warranted. We plan on future simulation studies to gain a quantitative understanding of classification performance with unlabeled data, including efficacy of finite sample statistical estimators in these contexts.

Again, a more detailed description of methodology and results is found in the peer-reviewed journal publication reproduced in **Appendix C**. [A.R. Jamieson, ML Giger, et.al. "Enhancement of breast CADx with unlabeled data," Medical Physics 37, 4155 (2010).]

3. Initial Investigation Parametric Deep Parametric Embeddings for CADx

The methods discussed above do not learn parametric mappings. In other words, a single data set may be dimension reduced to examine structure, but there is no generalization learned for future "out-of-sample" cases. This impedes practical application, including for breast CADx. Approximate methods have been suggested as a solution, but are not stable and prone to error. However, deep neural networks have been proposed for learning parametric embeddings.[11]

A deep neural network, or deep network, is a multi-layer neural network with more than one hidden unit layer. Deep networks are known to learn more efficiently and robustly represent (arbitrarily) complex functions than shallow networks (single hidden layer).[12] However, deep networks can be difficult to train due to the large number of weights and susceptibility to over-fitting. These problems are partially mitigated by a three step training procedure. First, the weights are initialized by "pre-training" each layer, one at a time, as a stacked Restricted Boltzmann Machine (RBM). Second, the separate layers are then connected to form a feed-forward neural network. Third, the network is fine-tuned using back-propagation minimizing a cost function**.** The same cost function as t-SNE can be chosen to help preserve local structure in the reduced embedding space. Supervised cost functions can also be used.[13] Provided training problems are overcome, deep parametric embeddings are perhaps the most compelling solution for understanding high-dimensional breast CADx feature spaces. We also explored deep parametric embeddings as another alternative for using unlabeled data.

KEY RESEARCH ACCOMPLISHMENTS

Year 1 (previously reported)

- Designed and executed Swift script enabled Grid computing work-flows, and managed to display clear proof of principle for large scale breast image analysis.

- Made effective use of the new 256 CPU SIRAF Shared Computing Facility at the University of Chicago Dept. of Radiology both as a test-bed for large scale parallel job submission and workflow management and to rapidly conduct new Multi-Modality Breast Image CADx research, culminating in a peer-reviewed journal article submission. Estimated Total CPU time used: ~100,000 to 300,000 hours.

- Investigated the use of cutting edge data-analysis/mining techniques as applied to Ultrasound, FFDM, and DCE-MRI Breast Image Feature Space Analysis for CADx , specifically, dimension reduction and data representation techniques (t-SNE and Laplacian Eigenmaps) for high dimensional data spaces. These methods allow for an alternative to traditional feature selection methods. Using the high-throughput cluster computing capabilities, performance metrics and intensive statistical cross-validation (0.632+ bootstrap and ROC analysis for AUC performance) were performed to gain understanding of the new techniques potential versus previous Breast CADx methodologies. Results indicate the ability to rival or exceed previous CADx performance.

- The dimensional reduction and data representation techniques also were shown to provide rich visualization output for human interpretation of the complex breast image feature space geometry.

- Additionally, the promising findings and have motivated a number of new research avenues. Most significantly, the incorporation and principled use of "unlabeled" (truth-unknown/non-biopsy proven) image data for the training of CADx algorithms. Specifically, the unsupervised dimension reduction techniques can use the feature space geometric structure to help regularize algorithmic training.

Year 2

- Continued use of 256 CPU SIRAF Shared Computing Facility at the University of Chicago Dept. of Radiology to rapidly complete experiments. Estimated Total CPU time used: >50,000 hours.

- Development of novel breast CADx schemes for incorporating unlabeled data into the algorithm training.

- Experimental evaluation of new breast CADx schemes capable of incorporating unlabeled data using previously acquired breast image feature data.

- Experimental results detected statistically significant improvements when using unlabeled data.

- Explored the use of parametric embedding algorithms for more practical application of dimension reduction in breast CADx

REPORTABLE OUTCOMES

Peer-reviewed Journal Publications

A.R. Jamieson, M. L. Giger, et. al. "Exploring Non-Linear Feature Space Dimension Reduction and Data Representation in Breast CADx with Laplacian Eigenmaps and t-SNE," Medical Physics. 37, 339 (2009).

A.R. Jamieson , M.L. Giger et. al. "Enhancement of breast CADx with unlabeled data," Medical Physics 37, 4155 (2010).

Conference Presentations and Abstracts

- **A.R. Jamieson**, M L Giger, M Wilde, L Pesce, I Foster, "Grid-Computing for Optimization of CAD," *Poster,* 50th Assembly and Annual Meeting of American Association of Physicist in Medicine (AAPM), Houston, Illinois, USA, July 2008

- **A.R. Jamieson**, ML Giger, L. Pesce "Regularized Training of CADx Algorithms with Unlabeled Data Using Dimension Reduction Techniques," A*ccepted talk.* 95nd Assembly and Annual Meeting of Radiological Society of North America, Chicago, Illinois, USA, December 2009.

- **A.R. Jamieson,** M L Giger, et. al. "Exploring Non-Linear Feature Space Dimension Reduction and Data Representation in Breast CADx", A*ccepted talk.* 51st Assembly and Annual Meeting of American Association of Physicist in Medicine (AAPM) Anaheim CA, USA, July 2009

- **A.R. Jamieson,** R. Alam, M.L. Giger. "Exploring Deep Parametric Embeddings for Breast CADx." *Accepted talk.* SPIE Medical Imaging 2011, Lake Buena Vista, Florida, USA

Invited/Misc. Journal Articles

A.R. Jamieson , M.L. Giger et. al. "Enhancement of breast CADx with unlabeled data," selected for the August 2010 issue of Virtual Journal of Biological Physics Research.

CONCLUSIONS

Overall, we are pleased to report successful research progress during the second year of funding. This productive year culminated in the publication of a journal article (attached for reference in **Appendix C**) summarizing our findings, and acceptance of an oral presentation and conference proceedings in 2011. The primary focus of this past year's research has been to investigate the use of unlabeled data for improving breast image CADx performance. Our results suggest strong evidence for the benefit of using unlabeled data. This is an important development for future breast image CADx research, including future large-scale Grid-enabled analysis approaches, since collecting labeled clinical image data information can be resource expensive, and limiting constraint. We expect, as digital imaging continues grow, more imaging data will become available, most of which will be unlabeled (histo-pathology unknown). Looking to future research, we will develop a more quantitative and theoretical understanding of classification performance when using unlabeled data via simulations studies. Additionally, we are focused on refining these techniques (e.g., such as parametric embeddings) to better prepare for actual future clinical application.

REFERENCES

1. Jemal, A., Siegel, R., Xu, J. & Ward, E. Cancer statistics, 2010. *CA Cancer J Clin* **60**, 277-300 (2010).
2. Tabar, L. & Dean, P. Thirty years of experience with mammography screening: a new approach to the diagnosis and treatment of breast cancer. *Breast Cancer Research* **10**, S3 (2008).
3. Tabar, L., Tot, T. & Dean, P.B. *Breast Cancer - The Art and Science of Early Detection with Mammography: Perception, Interpretation, Histopathologic Correlation*. (Thieme: 2004).
4. Jamieson, A.R. et al. Exploring nonlinear feature space dimension reduction and data representation in breast CADx with Laplacian eigenmaps and t-SNE. *Med. Phys.* **37**, 339-351 (2010).
5. Belkin, M., Niyogi, P. & Sindhwani, V. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *J. Mach. Learn. Res.* **7**, 2399-2434 (2006).
6. Jamieson, A.R., Giger, M.L., Drukker, K. & Pesce, L.L. Enhancement of breast CADx with unlabeled data. *Med Phys* **37**, 4155-4172 (2010).
7. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2605, 2579 (2008).
8. Belkin, M. & Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.* **15**, 1373-1396 (2003).
9. Sidak, Z. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *J. Am. Stat. Assoc.* **62**, 626-633 (1967).
10. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **6**, 65-70 (1979).
11. van der Maaten, L. Learning a Parametric Embedding by Preserving Local Structure. *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics* 384-391
12. Bengio, Y. *Learning Deep Architectures for AI*. (Now Publishers Inc: 2009).
13. Min, R., van der Maaten, L., Yuan, Z., Bonner, A. & Zhang, Z. Deep Supervised t-Distributed Embedding. *Proceedings of the International Conference on Machine Learning (ICML)* (2010).

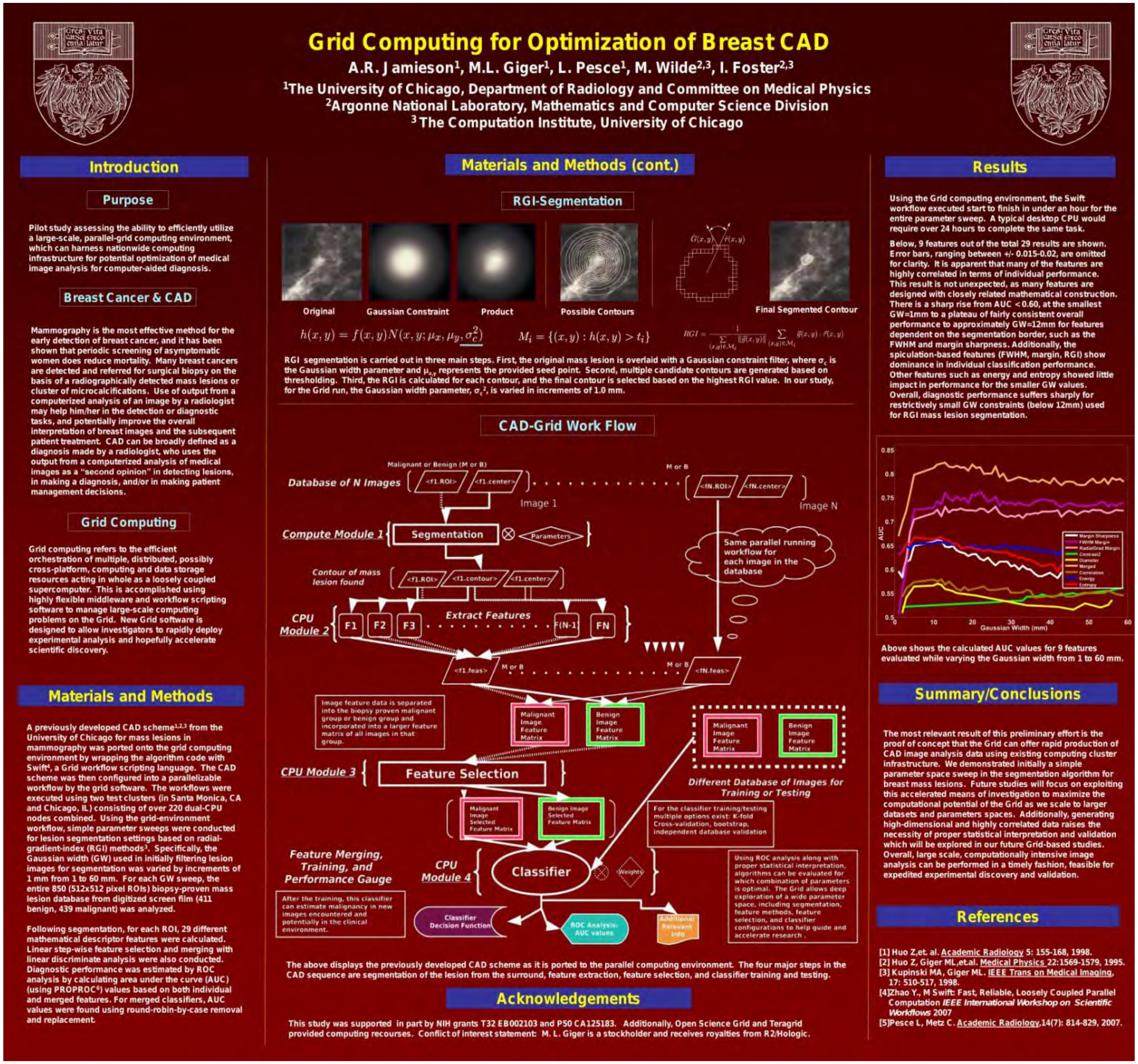
APPENDICES

- Appendix A: *POSTER*. **A.R**. **Jamieson**, M L Giger, M Wilde, L Pesce, I Foster, “Grid-Computing for Optimization of CAD,” *Poster,* 50th Assembly and Annual Meeting of American Association of Physicist in Medicine, Houston, Illinois, USA, July 2008

- Appendix B: *PAPER.* **A.R. Jamieson**, M. L. Giger, et. al. “Exploring Non-Linear Feature Space Dimension Reduction and Data Representation in Breast CADx with Laplacian Eigenmaps and t-SNE”, Medical Physics, 37, 339 (2010).

- Appendix C: *PAPER.* **A.R. Jamieson**, ML Giger, et.al. "Enhancement of breast CADx with unlabeled data," Medical Physics 37, 4155 (2010).

APPENDIX A



Grid Computing for Optimization of Breast CAD

A.R. Jamieson[1], M.L. Giger[1], L. Pesce[1], M. Wilde[2,3], I. Foster[2,3]
[1]The University of Chicago, Department of Radiology and Committee on Medical Physics
[2]Argonne National Laboratory, Mathematics and Computer Science Division
[3]The Computation Institute, University of Chicago

Introduction

Purpose

Pilot study assessing the ability to efficiently utilize a large-scale, parallel-grid computing environment, which can harness nationwide computing infrastructure for potential optimization of medical image analysis for computer-aided diagnosis.

Breast Cancer & CAD

Mammography is the most effective method for the early detection of breast cancer, and it has been shown that periodic screening of asymptomatic women does reduce mortality. Many breast cancers are detected and referred for surgical biopsy on the basis of a radiographically detected mass lesions or cluster of microcalcifications. Use of output from a computerized analysis of an image by a radiologist may help him/her in the detection or diagnostic tasks, and potentially improve the overall interpretation of breast images and the subsequent patient treatment. CAD can be broadly defined as a diagnosis made by a radiologist, who uses the output from a computerized analysis of medical images as a "second opinion" in detecting lesions, in making a diagnosis, and/or in making patient management decisions.

Grid Computing

Grid computing refers to the efficient orchestration of multiple, distributed, possibly cross-platform, computing and data storage resources acting in whole as a loosely coupled supercomputer. This is accomplished using highly flexible middleware and workflow scripting software to manage large-scale computing problems on the Grid. New Grid software is designed to allow investigators to rapidly deploy experimental analysis and hopefully accelerate scientific discovery.

Materials and Methods

A previously developed CAD scheme[1,2,3] from the University of Chicago for mass lesions in mammography was ported onto the grid computing environment by wrapping the algorithm code with Swift[4], a Grid workflow scripting language. The CAD scheme was then configured into a parallelizable workflow by the grid software. The workflows were executed using two test clusters (in Santa Monica, CA and Chicago, IL) consisting of over 220 dual-CPU nodes combined. Using the grid-environment workflow, simple parameter sweeps were conducted for lesion segmentation settings based on radial-gradient-index (RGI) methods[3]. Specifically, the Gaussian width (GW) used in initially filtering lesion images for segmentation was varied by increments of 1 mm from 1 to 60 mm. For each GW sweep, the entire 850 (512x512 pixel ROIs) biopsy-proven mass lesion database from digitized screen film (411 benign, 439 malignant) was analyzed.

Following segmentation, for each ROI, 29 different mathematical descriptor features were calculated. Linear step-wise feature selection and merging with linear discriminate analysis were also conducted. Diagnostic performance was estimated by ROC analysis by calculating area under the curve (AUC) (using PROPROC[5]) values based on both individual and merged features. For merged classifiers, AUC values were found using round-robin-by-case removal and replacement.

Materials and Methods (cont.)

RGI-Segmentation

Original — Gaussian Constraint — Product — Possible Contours — Final Segmented Contour

$$h(x,y) = f(x,y)N(x,y;\mu_x,\mu_y,\underline{\sigma_C^2})$$

$$M_i = \{(x,y) : h(x,y) > t_i\}$$

$$RGI = \frac{1}{\sum_{(x,y)\in M_i} \|\vec{g}(x,y)\|} \sum_{(x,y)\in M_i} \vec{g}(x,y)\cdot\hat{r}(x,y)$$

RGI segmentation is carried out in three main steps. First, the original mass lesion is overlaid with a Gaussian constraint filter, where σ_c is the Gaussian width parameter and $\mu_{x,y}$ represents the provided seed point. Second, multiple candidate contours are generated based on thresholding. Third, the RGI is calculated for each contour, and the final contour is selected based on the highest RGI value. In our study, for the Grid run, the Gaussian width parameter, σ_c^2, is varied in increments of 1.0 mm.

CAD-Grid Work Flow

Malignant or Benign (M or B)

Database of N Images {<f1.ROI> <f1.center>} Image 1 . . . . M or B {<fN.ROI> <fN.center>} Image N

Compute Module 1 {Segmentation ⊗ Parameters}

Contour of mass lesion found {<f1.ROI> <f1.contour> <f1.center>}

Same parallel running workflow for each image in the database

CPU Module 2 {F1 F2 F3 . . . . F(N-1) FN} Extract Features

<f1.feas> M or B . . . . M or B <fN.feas>

Image feature data is separated into the biopsy proven malignant group or benign group and incorporated into a larger feature matrix of all images in that group.

Malignant Image Feature Matrix — Benign Image Feature Matrix

CPU Module 3 {Feature Selection}

Malignant Image Selected Feature Matrix — Benign Image Selected Feature Matrix

Malignant Image Feature Matrix — Benign Image Feature Matrix

Different Database of Images for Training or Testing

For the classifier training/testing multiple options exist: K-fold Cross-validation, bootstrap, independent database validation

Feature Merging, Training, and Performance Gauge

CPU Module 4 {Classifier ⊗ Weights}

After the training, this classifier can estimate malignancy in new images encountered and potentially in the clinical environment.

Classifier Decision Function — ROC Analysis: AUC values — Additional Relevant Info

Using ROC analysis along with proper statistical interpretation, algorithms can be evaluated for which combination of parameters is optimal. The Grid allows deep exploration of a wide parameter space, including segmentation, feature methods, feature selection, and classifier configurations to help guide and accelerate research .

The above displays the previously developed CAD scheme as it is ported to the parallel computing environment. The four major steps in the CAD sequence are segmentation of the lesion from the surround, feature extraction, feature selection, and classifier training and testing.

Acknowledgements

This study was supported in part by NIH grants T32 EB002103 and P50 CA125183. Additionally, Open Science Grid and Teragrid provided computing recourses. Conflict of interest statement: M. L. Giger is a stockholder and receives royalties from R2/Hologic.

Results

Using the Grid computing environment, the Swift workflow executed start to finish in under an hour for the entire parameter sweep. A typical desktop CPU would require over 24 hours to complete the same task.

Below, 9 features out of the total 29 results are shown. Error bars, ranging between +/- 0.015-0.02, are omitted for clarity. It is apparent that many of the features are highly correlated in terms of individual performance. This result is not unexpected, as many features are designed with closely related mathematical construction. There is a sharp rise from AUC < 0.60, at the smallest GW=1mm to a plateau of fairly consistent overall performance to approximately GW=12mm for features dependent on the segmentation border, such as the FWHM and margin sharpness. Additionally, the spiculation-based features (FWHM, margin, RGI) show dominance in individual classification performance. Other features such as energy and entropy showed little impact in performance for the smaller GW values. Overall, diagnostic performance suffers sharply for restrictively small GW constraints (below 12mm) used for RGI mass lesion segmentation.

Above shows the calculated AUC values for 9 features evaluated while varying the Gaussian width from 1 to 60 mm.

Summary/Conclusions

The most relevant result of this preliminary effort is the proof of concept that the Grid can offer rapid production of CAD image analysis data using existing computing cluster infrastructure. We demonstrated initially a simple parameter space sweep in the segmentation algorithm for breast mass lesions. Future studies will focus on exploiting this accelerated means of investigation to maximize the computational potential of the Grid as we scale to larger datasets and parameters spaces. Additionally, generating high-dimensional and highly correlated data raises the necessity of proper statistical interpretation and validation which will be explored in our future Grid-based studies. Overall, large scale, computationally intensive image analysis can be performed in a timely fashion, feasible for expedited experimental discovery and validation.

References

[1] Huo Z,et. al. Academic Radiology 5: 155-168, 1998.
[2] Huo Z, Giger ML,et.al. Medical Physics 22:1569-1579, 1995.
[3] Kupinski MA, Giger ML. IEEE Trans on Medical Imaging, 17: 510-517, 1998.
[4]Zhao Y., M Swift: Fast, Reliable, Loosely Coupled Parallel Computation *IEEE International Workshop on Scientific Workflows* 2007
[5]Pesce L, Metz C. Academic Radiology,14(7): 814-829, 2007.

APPENDIX B

Manuscript in Publication: *Med. Phys.* 37, 339-351 (2010)

Title:

Exploring Non-Linear Feature Space Dimension Reduction and Data Representation in Breast CADx with Laplacian Eigenmaps and t-SNE

Authors: Andrew R. Jamieson, Maryellen L. Giger, Karen Drukker, Hui Li, Yading Yuan, and Neha Bhooshan
Department of Radiology, University of Chicago, Chicago, Illinois 60637

Abstract:

In this preliminary study, recently developed unsupervised non-linear dimension reduction (DR) and data representation techniques were applied to computer-extracted breast lesion feature spaces across three separate imaging modalities: ultrasound (US) with 1126 cases, dynamic contrast enhanced-magnetic resonance imaging (DCE-MRI) with 356 cases, and full-field digital mammography (FFDM) with 245 cases. Two methods for non-linear DR were explored: Laplacian Eigenmaps of Belkin and Niyogi,[1] and t-distributed stochastic neighbor embedding (t-SNE) of van der Maaten and Hinton.[2] These methods attempt to map originally high-dimensional feature spaces to more human interpretable lower-dimensional spaces while preserving both local and global information. The properties of these methods as applied to breast computer-aided diagnosis (CADx) were evaluated in the context of malignancy classification performance as well as in the visual inspection of the sparseness within the two- and three-dimensional mappings. Classification performance was estimated by using the reduced dimension mapped feature output as input into both linear and non-linear classifiers: Markov Chain Monte Carlo based Bayesian artificial neural network (MCMC-BANN) and linear discriminate analysis (LDA). The new techniques were compared to previously developed breast CADx methodologies, including Automatic Relevance Determination (ARD) and linear step-wise (LSW) feature selection, as well as a linear DR method based on Principal Component Analysis (PCA). Using ROC analysis and 0.632+ bootstrap validation, 95% empirical confidence intervals were computed for the each classifier's AUC performance. Results: In the large US dataset, sample high performance results include, $AUC_{0.632+}$ = 0.88 with 95% empirical bootstrap interval [0.787;0.895] for 13 ARD selected features and $AUC_{0.632+}$ = 0.87 with interval [0.817;0.906] for 4 LSW selected features compared to 4D t-SNE mapping (from the original 81D feature space) giving $AUC_{0.632+}$ = 0.90 with interval [0.847;0.919], all using the MCMC-BANN. Conclusions: Preliminary results appear to indicate capability for the new methods to match or exceed classification performance of current advanced breast lesion CADx algorithms. While not appropriate as a complete replacement of feature selection in CADx problems, DR techniques offer a complementary approach which can aid elucidation of additional properties associated with the data. Specifically, the new techniques were shown to possess the added benefit of delivering sparse lower-dimensional representations for visual interpretation, revealing intricate data structure of the feature space.

Keywords: non-linear dimension reduction, computer-aided diagnosis, breast cancer, Laplacian Eigenmaps, t-SNE

I. Introduction

Radiologic image interpretation is a complex task. A radiologist's expertise, developed only with exhaustive training and experience, rests in their ability for extracting and meaningfully synthesizing relevant information from a medical image. However, even under idealized image acquisition conditions, precise conclusions may not be possible for certain radiologic tasks. Thus, computer aided diagnosis (CADx) systems have been introduced in a number of contexts in an attempt to assist human interpretation of medical images.[3] A relatively well-developed clinical application for which computerized efforts in radiological image analysis have been studied is the use of CAD in the task of detecting and diagnosing breast cancer.[4-10] Similar to the radiologist’s task, a computer algorithm is designed to make use of the highly complicated breast image input data, attempting to intelligently reduce image information into more interpretable and ultimately clinically-actionable output structures, such as an estimate of the probability of malignancy. Understanding how to optimally make use of the enormity of the initial image information input and best arrive at the succinct conceptual notion of “diagnosis” is a formidable challenge. Although there may be any number of various operations/transformations involved in arriving at this high-level end output, whether in the human brain or *in silico*, two common critical pursuits are proper data representation and reduction. The current study aims to explore the potential enhancements offered to breast mass lesion CADx algorithms through the application of two recently-developed dimensionality reduction and data representation techniques, Laplacian Eigenmaps and t-distributed stochastic neighbor embedding (t-SNE).[1,2]

II. Background

II.A. Current CADx Feature Representation

Restricted by limited sample datasets, computational power, and lack of complete theoretical formalism, image-based pattern recognition and classification techniques often tackle the objective task at hand by substantially simplifying the problem. Traditionally, breast CADx systems employ a two pronged approach, first, image pre-processing and feature extraction, and second, classification in the feature space, either by unsupervised methods, supervised methods, or both. A review of past and present CADx methods employed can be found in referenced articles referenced.[3,11] Often, instead of attempting to make use of the complete image[12], CADx typically condenses image information down to a vector of numerical values, each representative of some attribute of the image or lesion present in the image. One can consider this first data reduction step as “perceptual” processing, meaning that at this stage the algorithm’s goal is to isolate and "perceive" only the most relevant components of the original image that will contribute towards distinguishing between the target classes (e.g., malignant or benign). One of the steps in eliminating unnecessary image information is lesion margin segmentation. [5,13] Typically, features, such as those extracted from the segmented lesion, are heuristic in nature and mimic important human identified aspects of the lesion. However more mathematical and abstract feature quantities may also be calculated that may represent information visually imperceptible to the unaided eye. While the use of data from a segmented lesion introduces bias into the algorithm's task as a whole, this "informed" bias allows for the efficient removal of much unnecessary image data, for instance normal background breast tissue. From here the second main component of the CADx algorithm falls usually into the context of the well-formalized canonical problem found in statistical pattern recognition for classification[14,15].

After the first CADx phase of feature extraction, each high-dimensional image in the sample set is now reduced to a single vector in a lower-dimensional feature space. However, due to the finite size of image sample data, if too many features are examined simultaneously, regions containing a low density of points in the feature space will exist, resulting in statistically inconclusive classification ability. This dilemma is affectionately termed the "curse of dimensionality."[16] Thus, a further reduction in the full feature space is required for a practically useful data representation. This aspect is a major concern of the second component of traditional CADx schemes, and is succinctly known as "feature selection". Much literature has been generated on this subject matter in the explicit context of improving CADx performance [17-19] Some CADx schemes may employ only 4-5 features maximum, in which case, feature selection may not be necessary, since the dataset sample size, even for relatively smaller sizes, may be sufficiently large to avoid over-training classifiers. However, it is reasonable to imagine CADx researchers interested in testing hundreds of potential features. In either case, when appropriately coupled with a well-regularized supervised classification method, the ultimate objective of features selection is to discover the "optimal" data representation, or sub-set of features for robustly maximizing the desired diagnostic task performance. That is, the method attempts both to mimic and to maximize the theoretical upper bound or ideal observer performance possible over the sampled joint probability distribution of the selected features. While this step is critical, finding such a sub-set is non-trivial and may also be highly dependent on the specific characteristics of the sample data. Developed techniques in feature selection for CADx range from simpler linear methods, such as those based on linear discriminate analysis (LDA), to non-linear and more sophisticated Bayesian-based, such as the use of Bayesian Artificial Neural Networks (BANN) and Automatic Relevance Determination (ARD), to random-search stochastic methods such as genetic algorithms as well as information theoretic techniques[17,19-21].

The most striking quality of the methods mentioned above, in the context of CADx, is that during feature selection, some features are completely removed from the final classification scheme, and hence image information is either explicitly or implicitly discarded altogether. However, while removing out all the information associated with a specific feature not selected, by selecting a smaller sub-set of individual features, what is gained is greater immediate human interpretability. Specifically, the isolated groups of features may have clear physical or radiological meanings and thus may be of interest to investigators or radiologists for understanding how these characteristics relate to the ability to distinguish class categories (malignant, benign, cyst, etc..). To this end, in order to interpret the nature of the feature space and attempt to identify characteristic trends, one may visually inspect plots displaying single features or attempt to capture synergistic qualities between two or three features simultaneously. Above three dimensions, as it becomes non-trivial to interpret the structure of the feature space, often instead, the use of a metrics such at the ROC curve and/or AUC based on output from the decision variable of a trained merged feature classifier are used to interrogate the quality of the higher dimensional feature spaces.

As such, beyond identifying which feature or features appear to hold classification utility, current CADx methods offer little theoretical/formal guidance in a recovering understanding of the inherent data structure represented by the higher dimensional feature spaces.

II.B. Proposed Feature Space Representation and Reduction for CADx

Due in part to the ever-growing demand of data driven science, in recent years much interest has emerged in developing techniques for discovering efficient representations of large-scale complex data.[22] Conceptually the goal is to discover the intrinsic structure of the data and adequately express this information in a lower dimensional representation. Classically, the problem of dimension reduction(DR) and data representation has been approached by applying linear transformations such as the well-known Principal Component Analysis (PCA) or more general Singular Value Decomposition (SVD). [23,24] Interestingly, despite PCA's age, only recently has this method been considered for the specific application to CADx feature space reduction.[25] In this particular breast ultrasound study, while no significant boosts in lesion classification performance were discovered, PCA was found to be a suitable substitute in place of more computationally intensive and cumbersome feature selection methods.[25] This efficient lower dimensional PCA data representation, i.e. linear combinations of the original features accounting for the maximum global variance decomposition in the data, proved capable of capturing sufficient information for robust classification. However, PCA is not capable of representing higher order, non-linear, local structure in the data.

The goal of recently proposed non-linear data reduction and representation methods focuses on this very problem.[1,2] The present methods of interest to this study, Laplacian Eigenmaps and t-Distributed Stochastic Neighbor Embedding (t-SNE), offer two distinct approaches for explicitly addressing the challenge of capturing and efficiently representing the properties of the low dimensional manifold on which the original high-dimensional data may lie. Previous studies have investigated other non-linear DR techniques, including self-organizing maps (SOMs) and graph embedding, for breast cancer in the context of biomedical image signal processing[26,27], as well as for a breast cancer BIRADs database clustering.[28] To our knowledge the relationship between breast CADx performance and these non-linear feature space DR and representation have yet to be properly investigated. These new techniques may contribute two key enhancements to current CADx schemes.

1. A principled alternative to feature selection. Both methods explicitly attempt to preserve as much structure in the original feature space as possible, and thus require no need to assumingly force exclusion of features from the original set, and hence unnecessary loss of image information.

2. A more natural and sparse data representation that immediately lends itself to generating human-interpretable visualizations of the inherent structures present in the high-dimensional feature data.

It is important to note that by employing DR on CADx feature spaces, one surrenders, to a varying extent, the ability to immediately interpret the physical meaning of the embedded representation. Yet, critically, this is a necessary and fundamental trade-off, as the conceptual focus is shifted to a more holistic approach, specifically, that of discovering an efficient lower dimensional representation of the intrinsic data structure. The core tenant of such an unsupervised approach is to limit assumptions imposed on the data. This major shift in philosophy regarding the original high dimensional feature space embodies the notion, “let the data speak for itself.” It seems reasonable to assume that if supervised classifiers are capable of uncovering sufficient data structure in the extracted feature space for producing adequate classification performance, then such principled local geometry preserving reduction mappings should reveal structural evidence corroborating such findings.

II.C. Outline of Evaluation for Proposed Methods

The primary objective of this study is to evaluate the classification performance characteristics of breast lesion CADx schemes employing the Laplacian Eigenmap or t-SNE DR techniques in place of previously developed feature-selection methods. Secondly, and more qualitatively, we aim to investigate and gain insight into the properties of sample visualizations representative of lower-dimensional feature space mappings of high-dimensional breast lesion feature data. Additionally, the feasibility and robustness of these non-linear reduction methods for CADx feature space reduction are tested across three separate imaging modalities: ultrasound (US), dynamic contrast enhanced MRI (DCE-MRI), and full-field digital mammography (FFDM), having case sets of 1126 case, 356 cases, and 245 cases, respectively.

III. Methods

III.A. Dataset

All data characterized in this study consists of clinical breast lesions presented in images acquired at the University of Chicago Medical Center. Lesions are labeled according to the truth known by biopsy or radiologic report and collected under HIPAA-compliant IRB protocols. Furthermore, the breast lesion feature datasets were generated from previously developed CADx algorithms at the University of Chicago. For a review of these techniques see Giger, Huo, Kupinski for X-ray mammography, Drukker, for US, and Chen for DCE-MRI. [4-11,29]

In each of the modalities, the lesion center is identified manually for the CADx algorithm, which then performs automated-seeded segmentation of the lesion margin followed by computerized feature extraction. Table 1 below summarizes the content of the respective imaging modality databases used, including the total number of initial lesion features extracted. Note that the mammographic imaging modality (FFDM) contains only two lesion class categories, malignant and benign. For ultrasound and DCE-MRI a more detailed sub-categorization is provided, including invasive carcinoma (IDC), ductal carcinoma *in situ* (DCIS), benign solid masses, and benign cystic masses. For clarity, this initial study only considers binary classification performance in the task of distinguishing between the more broad identity of malignant and benign (cancerous vs.

non-cancerous). However, during qualitative inspection of the dimension reduced mappings, it will be of interest to re-introduce these distinctions for visualization purposes.

Modality	**Total Number of Images**	**Number of Malignant Lesions**	**Number of Benign Lesions**	**Total Number of Lesion Features Calculated**
US	2956	158	968 (401 mass / 567 cystic)	81
DCE-MRI	356	223 (151 IDC / 72 DCIS)	133	31
FFDM	735	132	113	40

Table 1. Feature Database Characteristics.

Geometric, texture, and morphological features, such as margin sharpness, were extracted across all modalities. Also, the DCE-MRI dataset includes kinetic features, and the US features include those related to posterior acoustic behavior.[8,10] All raw extracted feature value datasets were normalized to zero mean and divided by the unit sample standard deviation. Due to page limitations, the details of each feature can be found in the referenced papers.[4-11,29]

III.B. Classifiers

In our evaluation of the new DR techniques, we chose two types of classifiers: a relatively simple linear discriminant analysis (LDA) classifier and a more sophisticated non-linear, Bayesian artificial neural network, classifier (BANN). [15] LDA is a well-known and commonly used linear classification method which will not be reviewed here, for reference and examples in breast lesion CADx see references. [4,30,31] The BANN, as the name suggests, follows the usual multi-layer-perception, neural network design, but additionally employs Bayesian theory as a means of classifier regularization [15,32]. The BANN has been shown to model the optimal ideal observer for classification given sufficient sample sizes as input for training.[33] The critical technical hurdle in implementing BANNs lies in accurately estimating posterior weight distributions, as analytical calculation is intractable. As such, either approximation or sampling based methods must be deployed in practice.[34] Markov Chain Monte Carlo (MCMC) sampling methods can be used to directly sample from the full posterior probability distribution.[32] We implemented a MCMC-BANN classifier using Nabney's *Netlab* package for MATLAB.[35] The following network architecture, *k*--(*k*+1)--1, was used. That is, *k* input layer nodes (one for each of the *k* selected features), a hidden layer with (*k* + 1) nodes, and a single output target as probability of malignancy. For each classifier trained, we generated at least 2000 MCMC samples of the weights' posterior probability distribution. The mean value of the classification prediction (probability of malignancy) output from each of the different 2000 weight samples was used to produce a single classification estimate for new test input cases.

III.C. Explicit Supervised Feature Selection Methods

Two previously developed feature selection methods are considered in this paper for comparison, and include linear step-wise and ARD feature selection. These methods are used to identify a specific set of features for input into the classifier.

Linear Stepwise Feature Selection

Linear step-wise feature selection (LSW-FS) relies on linear discriminant-based functions. Beginning with only a single selected feature, multiple combinations of features are considered one at a time, by exhaustively adding, retaining, or removing each subsequent feature to the potential set of selected features. For each new combination, a metric, the Wilks' lambda is calculated and a selection criterion based on F-statistics is used.[17] The "F-to-enter" and "F-to-remove" used in this study were automatically adjusted to allow for the specified number of features desired for US, DCE-MRI, and FFDM feature selection. For examples of LSW-FS use in breast CADx references are provided.[17,25,30]

Automatic Relevance Determination

A consequence of the BANNs is the possibility for joint feature selection and classification using Automatic Relevance Determination (ARD).[15,32,34,35] ARD works by placing Bayesian hyper-priors, also known as hierarchical priors, over the initial prior distributions already imposed on the network weights connected to the input nodes. The "relevant" features are then discovered as estimates for the hyper parameters, which characterize the prior distributions over the respective input layer weights, are updated via Gibbs sampling giving the posterior hyper-parameter estimate. The magnitudes of the final, converged upon hyper-parameters are then used to indicate the relative utility of the respective feature input layer weights towards accomplishing the classification task. Thus, by way of the Bayesian regularization, ARD allows for one-shot feature selection and classifier design. Furthermore, a key advantage of ARD feature selection is its ability to identify important non-linear features coupled to the classification objective, due to the inherent non-linear nature of BANN.[19] Due to these qualities, ARD-MCMC-BANN classifiers were also included for comparison in our study.

In this study we extend MCMC-BANN to incorporate ARD following the implementation of Nabney. [35] This methodology was previously investigated for breast feature selection and classification in DCE-MRI CADx.[19] In our study, 1000 samples were calculated for the hyper-parameters beginning with a gamma hyper-prior distribution of mean parameter value equal to 3 and a shape parameter equal to 4.

III.D. Unsupervised Dimension-Reduction Feature Mappings

In comparison to the supervised feature-selection methods, three unsupervised DR methods were evaluated here; the latter two non-linear methods are offered as a novel application to the field of breast image CADx. The general problem of dimensionality reduction can be described mathematically as: provided an initial set $x_1, \ldots, x_k$ of k points in R^l , discover a set $y_1, \ldots, y_k$ in R^m such that $\boldsymbol{y}_i$ sufficiently describes or "represents" the qualities of interest found in the original set $\boldsymbol{x}_i$. In the context of breast lesion CADx feature extraction, the ideally lower dimensional mappings should aim to preserve and represent as much relevant structural information towards the task of malignancy estimation. It should be noted that DR still requires, in some sense, "feature selection," meaning, one must specify the number of mapped dimensions to retain for the subsequent classification step. Ideally, methods designed to estimate intrinsic dimensionality of the data structure could be used to direct this choice.[36] However, proper evaluation of the integrity of such methods in this context is beyond the scope of this research effort. Thus, in approaching the problem from a more naïve perspective, as done here, focus is centered on gaining a general intuition for the overall major trends encountered.

Linear Feature Reduction: PCA

Mathematically, PCA is linear transformation which maps the original feature space onto new orthogonal coordinates. The new coordinates, or principal components (PC), represent ordered orthogonal data projections capturing the maximum variance possible, with the first PC corresponding to the highest global variance.[23,24] Drukker, et al. used PCA as an alternative to feature selection for breast US CADx.[25]

Non-linear Feature Dimension Reduction

As discussed in the introduction and background sections, the following two recently proposed DR and data representation methods are non-linear in nature and specifically designed to address the problem of local data structure preservation. Laplacian Eigenmaps and t-SNE offer highly distinct solutions to this problem.

i. Laplacian Eigenmaps

Drawing on familiar concepts found in spectral graph theory, Laplacian Eigenmaps, proposed by Belkin and Niyogi in 2002, use the notion of a graph Laplacian applied to a weighted neighborhood adjacency graph containing the original data set information.[1] This weighted neighborhood graph is regarded geometrically as a manifold characterizing the structure of the data. The eigenvalues and eigenvectors are computed for the graph Laplacian which are in turn utilized for embedding a lower dimensional mapping representative of the original manifold. Acting as an approximation to the Laplace Beltrami operator, the weighted graph Laplacian transformation can be shown, in a certain sense, to optimally preserve local neighborhood information.[37] Thus, the feature data considered in the reduced dimensional space mapping is essentially a discrete approximate representation of the natural geometry of the original continuous manifold.

As Belkin and Niyogi note, the algorithm is relatively simple and straightforward to implement. Additionally, the algorithm is not computationally intensive. For our largest dataset the mappings were computed within a few seconds using MATLAB code. Algorithm details as well as explanation of necessary input parameters for the implementation used here are provided below in section VIII.A of the Appendix.

It is important to note that there is no theoretical justification for how to choose the needed parameters for the algorithm. Thus, an array of parameter choices was evaluated in this study. Lastly, parts of the MATLAB code, related only to the implementation of the Laplacian Eigenmap, were modified from the publically available dimension reduction toolbox provided by Laurens van der Maaten of Maasticht University.[38]

ii. t-Distributed Stochastic Neighbor Embedding (t-SNE*)*

The other non-linear mapping technique considered, t-Distributed Stochastic Neighbor Embedding (t-SNE) of van der Maaten and Hinton[2], approaches the dimension reduction and data reduction problem by employing entirely different mechanisms to the

Laplacian Eigenmaps. t-SNE attacks DR from a stochastic and probabilistic-based framework. While requiring orders of magnitude more computational effort, such statistically-oriented approaches, provided they are well-conditioned, may potentially offer greater flexibility in certain contexts due in part by the lessening of potentially restrictive theoretical mathematical formalism. For these reasons the t-SNE method was considered as an interesting comparison alongside the Laplacian Eigenmap.

t-SNE is an improved variation on the original stochastic neighbor embedding (SNE) of Hinton and Rowies.[39] The basic idea behind SNE is to minimize the difference between specially defined conditional probability distributions that represent similarities, calculated for the data points in both the high and low dimensional representations. In particular, SNE begins by first computing the conditional probability $p_{j|i}$ given by

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\| / 2\sigma_i{}^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\| / 2\sigma_i{}^2\right)} \quad \text{and} \quad q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)} \tag{1}$$

and $q_{j|i}$ in the lower dimensional space with $p_{i|i}$ and $q_{i|i}$ set to zero. These similarities express the probability that x_i (y_i) would select x_j (y_j) as its neighbor, resulting in high values for nearby points and lower values for distantly separated ones. The central assumption in SNE is that if the low-dimensional mapped points in Y space correctly model the similarity structure of its higher-dimensional counterparts in X, then the conditional probabilities will be equal. The summed Kullback-Leibler (KL) divergence is used to gauge how well $q_{j|i}$ models $p_{j|i}$. Using gradient descent methods, SNE minimizes a KL based cost function. Sampled points from an isotropic Gaussian with small variance centered at the origin are used to initialize the gradient decent. Updates are made to the mapped space Y for each iteration. Additionally, the parameter σ_i of eq (1) must be selected. σ_i is the variance in the Gaussian centered on the high dimensional point x_i. Because of the difficultly in determining if an optimal σ_i exists, a user defined property called perplexity is used to facilitate its selection, defined by $Perp(P_i)=2^{H(Pi)}$. Calculated in bits, $H(P_i)$ is the Shannon entropy over P_i

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i} \tag{2}$$

During SNE, a binary search is performed to find the value of σ_i that produces a P_i with the user specified perplexity. Suggested typical settings range between 5 and 50.[2]

t-SNE introduces two critical improvements to SNE.[2] First, the gradient as well as cost function optimization is simplified by using symmetrized conditional probabilities to define the joint probabilities on P and Q (e.g. $p_{ij} = (p_{j|i} + p_{i|j})/2n$) and the minimizing cost over a single KL divergence as opposed to a sum,

$$C = \sum_i KL(P_i \parallel Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \Rightarrow \quad C' = KL(P \parallel Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} . \tag{3}$$

Second, the distributional form of the low-dimensional joint probabilities is changed from a Gaussian, to the heavier tailed Student t-distribution with one degree of freedom. Roughly, this promotes a greater probability for moderately distanced data points in high dimensional space to be expressed by a larger distance in the low-dimensional map, thus more “faithfully” representing the original distance structure, and avoiding the “crowding problem." [2] The new q_{ij} is defined as

$$q_{ij} = \frac{\left(1+\left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{k \neq l}\left(1+\left\|y_k - y_l\right\|^2\right)^{-1}} \tag{4}$$

After incorporating the altered q_{ij}, the final gradient for the cost function is given by

$$\frac{\delta C}{\delta y_i} = 4\sum_i (p_{ij} - q_{ij})(y_i - y_j)(1+\left\|y_i - y_j\right\|^2)^{-1} \quad . \tag{5}$$

A step by step algorithm outline for t-SNE is provided in section VIII.B of the Appendix.

As recommended by Hinton and van der Maaten[2], PCA is first applied to the high-dimensional input data in order to expedite the computation of the pairwise distances. Lastly, as t-SNE was developed primarily for 2D and 3D data representation and visualization, it is important to note that the authors warn performance of t-SNE is not well understood for the general purpose of DR.[2] By applying t-SNE to the CADx feature reduction problem we hope to offer at least some empirical insight towards understanding its properties in such contexts. We used van der Maaten publicly available t-SNE MATLAB code and Intel processer optimized "fast_tsne" to generate the present data mappings[40].

III.E. Classifier Performance Estimation and Evaluation

The high-dimensional feature spaces DR methods were tested across all modalities for a range of lower target dimensions and user-defined algorithm parameters. We evaluated the classifier performance using the area under the Receiver Operating Curve ROC curve (AUC) via the non-parametric Wilcoxon-Mann-Whitney statistic, as calculated using the PROPROC software.[41-43] Statistical uncertainty in classification performance due to finite sample sizes was estimated by implementing 0.632+ bootstrapping methods for training and testing the classifiers.[44,31] Additionally, we computed the 95% empirical bootstrap confidence intervals on AUC values as estimated by no less than 500 bootstrap case set re-samplings. In all values reported, the sampling was conducting on a *by lesion* basis, as there may be multiple images associated with each unique lesion. In this regard, during classifier testing, the set of classifier outputs associated with a unique lesion were averaged to produce a single value. For the supervised feature selection methods (ARD and LSW), feature selection was conducted, up to the specified number of features, on each bootstrapped sample set. Notably, the more general MCMC-BANN was coupled with both the non-linear ARD and linear-based feature selection methods, while the linear LDA was only with the linear stepwise feature selection. As some of the calculations are computationally intensive, particularly the t-SNE mappings and MCMC-BANN training for the larger US data set, a 256-CPU shared computing resource cluster was employed to accomplish runs in a feasible time frame.

IV. Results

IV.A. Classification Performance.

MCMC-BANN and LDA classification performance is plotted as a function of the mapped or feature selected input space dimension for the three datasets, US, DCE-MRI, and FFDM, using the three different DR techniques, as well as the non-reduced selected features in Figure 1(a-f). Performance is characterized in terms of the 0.632+ bootstrapped AUC (left axis) and variability as gauged by the width of the empirical 95% bootstrap interval (right axis). The t-SNE perplexity was set to *Perp* = 30 and Laplacian Eigenmaps were generated with *Nearest Neighbor*=45 and *t*=1.0. Overall, the highest classification performance was attained by the largest sample-size US feature dataset with the DR-MCMC-BANN just slightly eclipsing the LDA, achieving approximately $AUC_{0.632+}$ ~ 0.90,while the smaller DCE-MRI and FFDM feature data produced peaks around $AUC_{0.632+}$ ~ 0.80. The variability in bootstrapped AUCs is also lowest for the large US dataset, hovering near ~ 0.07 as the number of inputs into the classifier is increased.

A few key observations can be made from the results regarding the use of DR. Primarily, the DR techniques, for both linear (PCA) and non-linear (t-SNE and Laplacian Eigenmaps), overall, appear to at least match, or and in some cases exceed, explicit feature selection classification $AUC_{0.632+}$ performance. This is most evident when compared to the the ARD-FS coupled with the MCMC-BANN performance across all three imaging modalities, Figures 1(a), 1(c), 1(e) (left axis). Specifically, in all cases the DR methods exhibited a more rapid rise to peak $AUC_{0.632+}$ performance and remained higher than the ARD-based feature selection for all dimension input sizes. Additionally, compared to the ARD feature selection approach, the DR methods produced less variability in the bootstrap AUC. Figures 1(a),1(c),1(e) (right axis) substantially highlight this phenomenon. In particular, for the US data, the ARD-FS variability, being greater that of than all the DR methods, clearly trends downward as more features are selected for input; gradually approaching the DR variability levels, yet usually remaining higher. By comparison, save for a slight increase at 1D, the DR variability is relatively consistent from 2D to 13D.

However, when coupled with the LSW feature selection, the MCMC-BANN produced more competitive results against the DR performance. For example, for this MRI data set, except for 10D and 11D, the LSW-MCMC-BANN edged above all the DR based methods. Likewise, the use of the LSW feature selection with the MCMC-BANN resulted in substantially reduced variation in classifier performance compared to the ARD-FS. The LSW-MCMC-BANN variation nearly matched the DR output for both the US and MRI across all input dimensions. For the FFDM data, except for 2D-5D, the LSW-MCMC-BANN held close to the DR variation level.

The less complex, yet more stable LDA classifier, Figures 1(b),1(d),1(f)(left axis), produced different characteristic results. In all cases the LSW-feature selection performance was initially higher, however, as the dimension input space was increased, the DR methods became comparable. Expectedly, when coupled with the linear LDA, the highly-non-linear stochastic based t-SNE DR consistently underperformed. Turning to variation for the LDA, Figure 1(b), 1(d), 1(f) (right axis), the LSW-FS again exhibited different behavior from ARD-FS, in that, except for the smaller-case-sized FFDM data, variability does not considerably fluctuate moving from 1D to 13D for both the LSW-FS and DR methods.

One manner by which to concisely analyze the performance characteristics of dimension-reduction/feature selection and classifiers designs for a particular dataset is to plot the bootstrap cross-validation AUC against the variability. An example is provided for the US feature dataset in Figure 2, with each point representing a different number of input dimensions. Data points located in the upper left corner indicate the most preferred performance qualities, i.e., higher classification performance and lower expected variability. Also provided in Figure 5, is a plot displaying classification results for both MCMC-BANN and LDA, in terms of the bootstrap AUC for the US data. Included within this plot are the empirical 95% confidence intervals to aid in gauging statistical significance for differences between estimated AUC values.

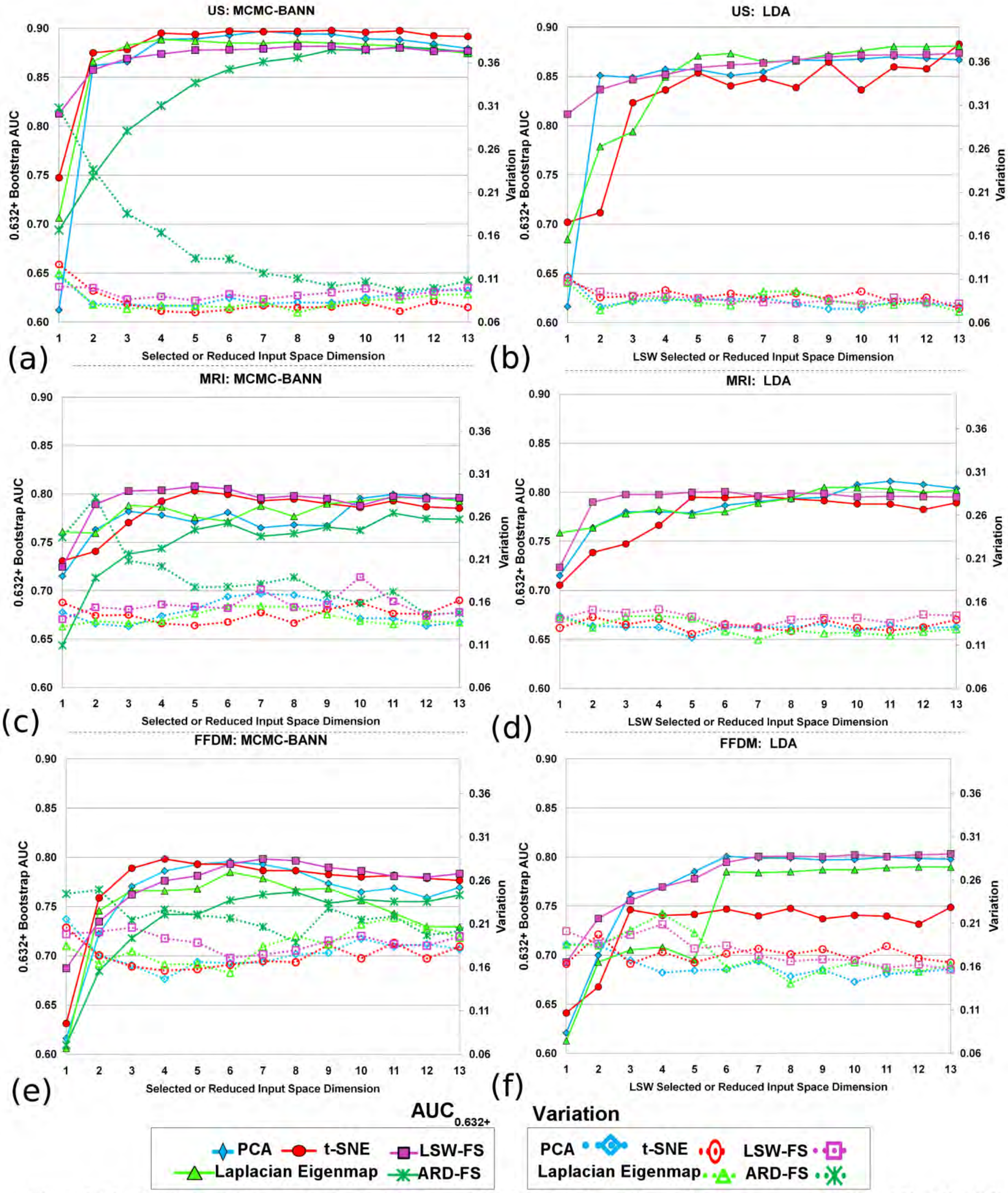


Figure 1. The 0.632+ bootstrap area under the ROC curve (AUC) (left axis) and the variation as measured by the width of the 95% empirical bootstrap confidence intervals (right axis) versus the selected feature {ARD,LSW} or reduced representation {PCA,t-SNE,Laplacian Eigenmap} classifier input space dimension. (a) MCMC-BANN, (b) LDA, classifier performance on the originally 81 dimensional US feature dataset. (c) MCMC-BANN, (d) LDA classifier performance on the originally 31 dimensional DCE-MRI feature dataset. (e) MCMC-BANN, (f) LDA classifier performance on the originally 40 dimensional FFDM feature dataset.

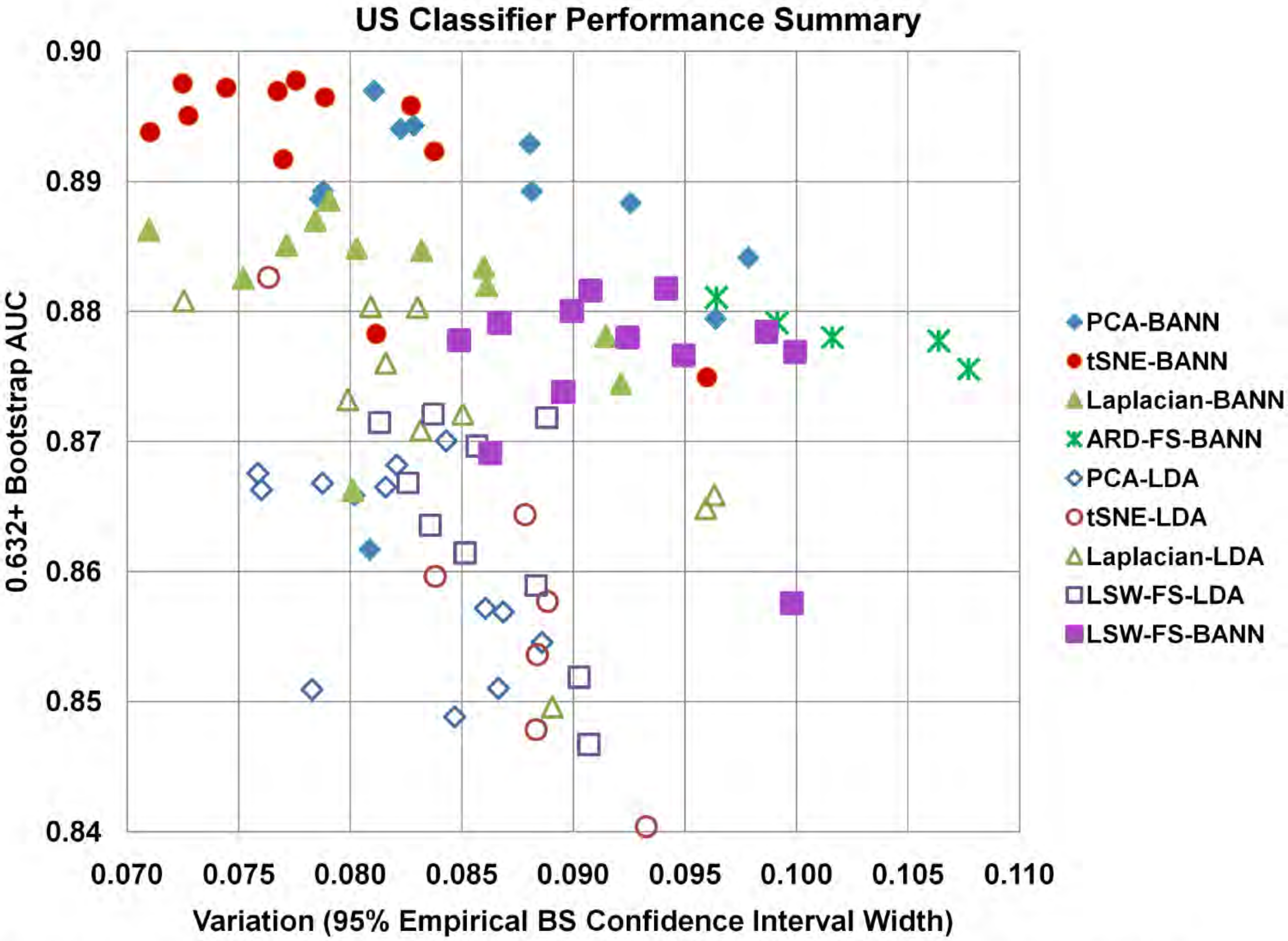


Figure 2. Summary of the classification performance on the 81 dimensional US feature dataset. The 0.632+ bootstrapped area under the ROC curve versus variability as gauged by the width of the 95% empirical bootstrap confidence intervals. Each point corresponds to a different input space dimension size. Points located in the upper left corner represent the highest expected AUC as well as least expected variation in performance due to sampling.

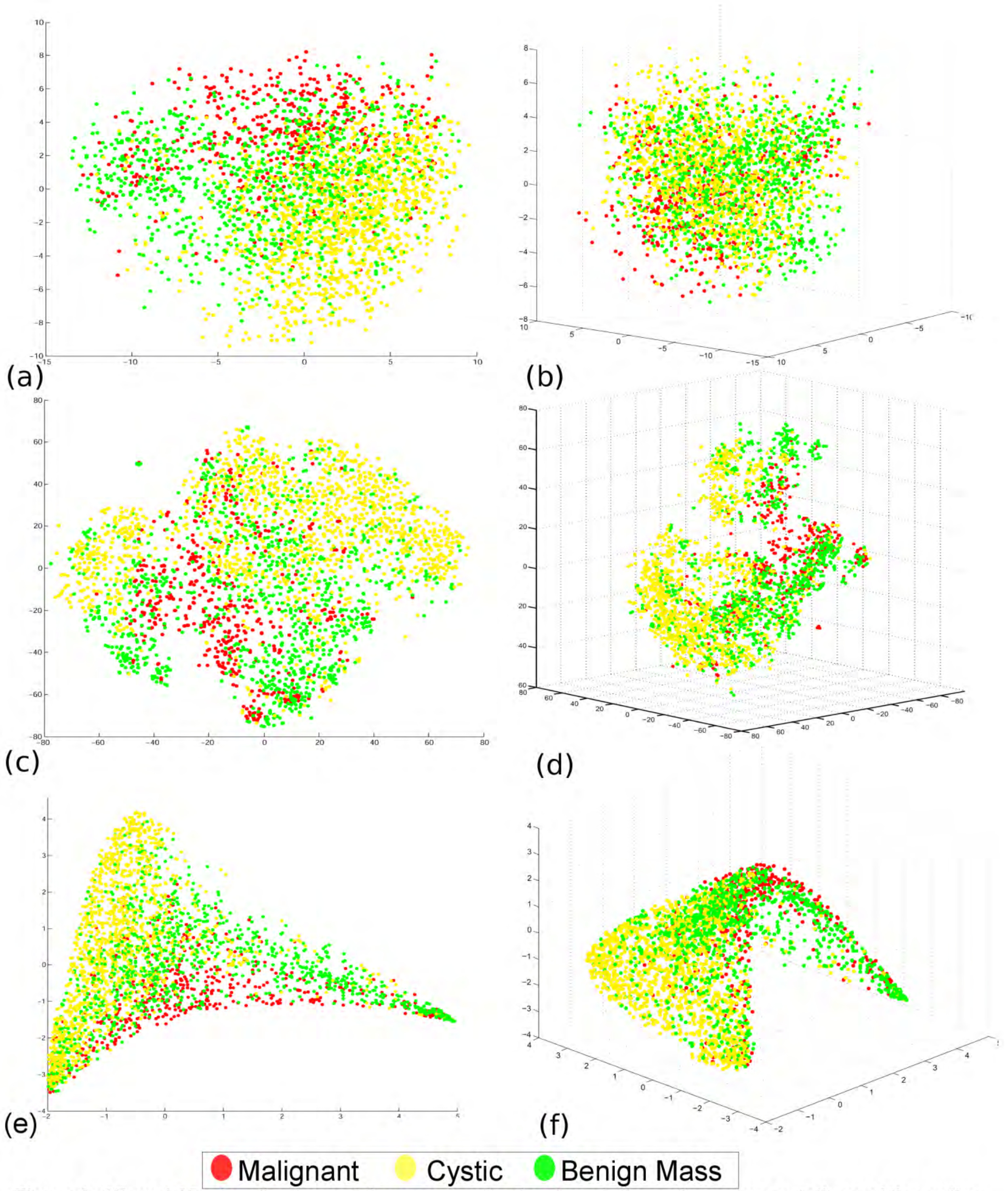


Figure 3. 2D and 3D visualizations of the unsupervised reduced dimension representations of the entire originally 81 dimensional breast lesion ultrasound feature dataset; green data points signifying benign lesions, red: malignant, and yellow: benign-cystic. (a) Visualization of linear reduction using PCA, first two principal components, (b) first three principal components, 3D PCA. (c) 2D and (d) 3D visualization of the non-linear reduction mapping using t-SNE. (e) 2D and (f) 3D visualization of the non-linear mapping using Laplacian Eigenmaps.

IV.B. 2D and 3D Visual Representations of Mappings

Due to the large sample size of the US feature data, a high density of points is produced (and hence the clearest delineation of structures) in the reduced dimension mapping representations. Figure 3(a-f) provides visual representations of the entire originally 81 dimensional US feature data mapped into 2D and 3D Euclidean space by the unsupervised PCA, t-SNE, and Laplacian Eigenmaps. The data points were subsequently colored to reflect the distribution of the lesions types (malignant tumor, benign lesion, cyst) with the reduced space.

Two key aspects are considered regarding the respective mappings: natural class separability and overall geometric traits characteristic of the represented structures, such as smoothness and sparsity. PCA is shown in Figures 3(a) and 3(b). Certain regions are potentially identifiable as being associated with a specific class (such as the dominance of cystic-benign points in the bottom right corner of the 2D plot), however, PCA generates a relatively homogeneous, nearly spherical distribution of points. Reflective of its mathematical basis, PCA representations provide primarily global information content, lacking the capability to represent rich local data structure. t-SNE generates a dramatically different type of low dimensional representation. As shown in figures 3(c) and 3(d), t-SNE produces a highly non-linear, jagged, and highly sparse data mapping. Many isolated "island-like" sub-groupings are identifiable in the t-SNE visual representations. As predicted by the high classification performance even for 2D and 3D, t-SNE manages to clearly capture inherent class structure associations. Lastly, the Laplacian Eigenmap, Figure 3(e) and 3(f), creates globally sparse, yet locally smooth representations. As captured by the figures, the distinctly triangular form in 2D is revealed as a projected aspect of a more complex, yet smoothly connected 3D geometric structure. As evident by upper "ridge" of malignant (red) lesion points and broad cystic (yellow) "fin" on the left, the Laplacian Eigenmap also manages to capture inherent class associations.

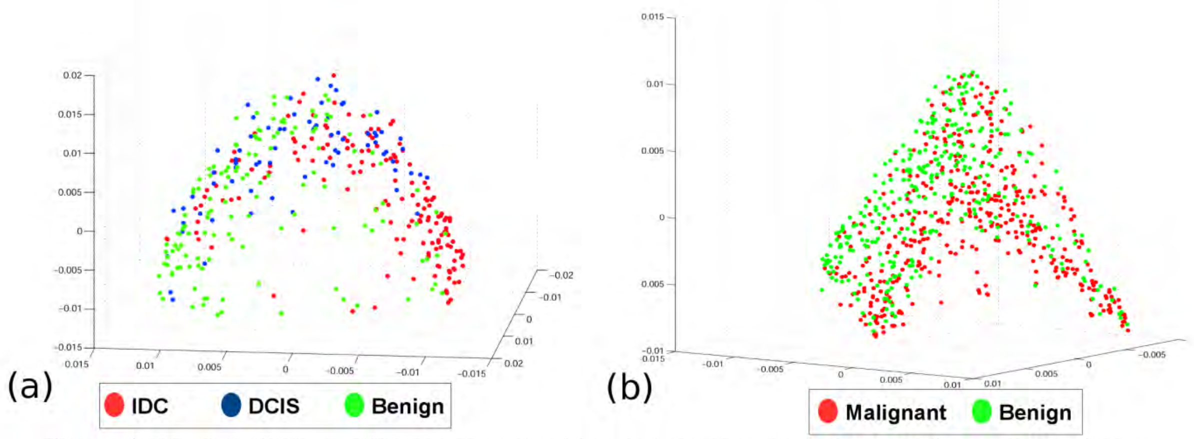


Figure 4. 3D visualization of the unsupervised local structure preserving non-linear dimension reduction representation using Laplacian Eigenmaps on breast lesion feature data. (a) 3D visualization of the entire originally 31 dimensional DCE-MRI feature data, green data points signify benign lesions, red: malignant-IDC, and blue: malignant-DCIS . (b) 3D visualization of the entire originally 40 dimensional FFDM feature data, green points for benign and red for malignant lesions.

The FFDM and DCE-MRI visual representations are noisier than the US due to the smaller sample size. A few examples are provided in Figure 4(a,b). The MRI dataset clearly exhibits a sparse arc-like geometric structure using the Laplacian Eigenmap. This structure seemingly separates the bulk of benign (green) lesions from the IDC (red) while dispersing the DCIS (blue) cases in between.

V. Discussion

V.A Dimension Reduction in CADx

Three major conclusions can be made regarding the use of DR techniques in breast CADx from this study. First, and most importantly, information critical for the classification of breast mass lesions contained within the original high-dimensional CADx feature vectors is not destroyed by applying the unsupervised, non-linear DR and representation techniques of t-SNE and Laplacian Eigenmaps. This observation is strongly supported by the robustness of the classification performance across the three different imaging modalities, US, DCE-MRI, and FFDM.

Second, according to the statistical re-sampling validation methods, the DR-based classification performance characteristics appear to potentially rival or in some cases exceed that of traditional feature-selection based techniques. Additionally, both the linear PCA and non-linear t-SNE and Laplacian Eigenmap methods often generated “tighter” 95% empirical bootstrap intervals, implying reduced variance in classifier output, as compared to the feature selection based approaches, especially ARD, see Figure (4). For instance, in the large US dataset, the performance for 13 ARD selected features was $AUC_{0.632+}$ = 0.88 with 95% empirical bootstrap interval [0.787;0.895] and for 4 LSW selected features was $AUC_{0.632+}$ = 0.87 with interval [0.817;0.906] compared to 4D t-SNE mapping (from the original 81D feature space) giving $AUC_{0.632+}$ = 0.90 with interval [0.847;0.919]. These findings imply that the generally non-linear manifold, on which US feature data exists, embedded in four dimensional Euclidean space can adequately represent the critical information for classification. These results build evidence for some potential benefits of employing the information-preserving, DR techniques in place of explicit feature selection, including the avoidance of the “curse of dimensionality”.

Third, the non-linear DR techniques generated visually-rich embedded mappings with a geometric structure that often presented sparse separation between class categories, as demonstrated in Figure 3(b): malignant, benign, cyst, and Figure 4(a): benign, DCIS, IDC. The natural class associations visible in the mappings are not totally unexpected since, as explored above, the classification performance results clearly demonstrate the reduced mapping’s capacity to retain sufficient information for class-discrimination. The large sample number of the US dataset provided the most vivid visualizations, highlighting both the geometric forms and sparse quality of the non-linear embeddings. Although PCA retained high supervised classification performance, unlike the non-linear Laplacian Eigenmaps and t-SNE embeddings, Figures 3(d),3(f), PCA is not capable of adequately representing the data’s inherent local structural properties, Figure 3(b), leading to less informative visualizations. Yet, the two non-linear methods offer distinct perspectives on the data structures. The Laplacian Eigenmap appears to perhaps frame the lesions in a more globally smooth context as evidenced by the gradual transitions between distant regions of the geometric form, whereas t-SNE creates many distinct jagged “islands” of clustered lesion points. These emergent characteristics reflect the theoretically motivated principles driving the respective non-linear DR algorithms.

V.B Reduction Method Parameters

We briefly explored the impact of the parameter selection towards performance and visual appearance. To our knowledge there is no principled way to optimally select a parameter configuration, thus we simply choose parameters that gave reasonable mappings as discernable in the 2D/3D representations. This is a problem in general for many unsupervised techniques. In fact, as t-SNE creators noted[2], the method was primarily considered for visualization purposes and not explicitly for DR beyond 3 dimensions. Performance of t-SNE is not well understood for the general purpose of DR and subsequent classification. Future work may be of interest to discover procedures for identifying "optimal" or "near-optimal" subsets of parameters for CADx or similar machine learning purposes.

V.C. Classifiers and Feature Selection

In considering classifier design, one desires to be "as simple as possible, but no simpler," meaning the most robust scheme in terms of both performance and stability (low variability in performance between different samples from the same underlying distribution), all while attempting to constrain the number of parameters, namely the input space dimension. Additionally, simpler models facilitate future repeatability with new contexts and datasets. The degree to which such pursuits are successful is dependent upon the interplay of the three main aspects affecting the performances of the classifiers including: sample size, data complexity, and model complexity/regularization. Naturally included within the scope of the model complexity/regularization is the choice of inputs to the classifier, whether in the form of DR mappings or a set of selected features, as this also critically influences ultimate classification capability. Ideally, any classifier's aim is to synthesize the information available from the input space in a complete and unbiased fashion towards accomplishing the decision task. In general, classification of new input based on finite training dataset is an "ill-posed" problem, and regardless of the sophistication of regularization employed, instability may persist. [15] For these reasons both the LDA and MCMC-BANN were investigated. By spanning over three different imaging modalities of varying data set size, using two different classifiers, and employing three different feature space approaches, all three of these key concepts (sample size, sample complexity, and model complexity) were touched upon in the course of this investigation.

For the relatively large US dataset, with 1126 unique lesions making up 2956 lesion images, some of the relative strengths associated with the more general, non-linear MCMC-BANN were particularly apparent. Specifically, the MCMC-BANN, when paired with either the DR techniques or LSW-FS was able to achieve high $\mathrm{AUC}_{0.632+}$ performance, even at low input space dimensions, as seen in Figure 1(a). This is in part due to the MCMC-BANN ability to generalize to any target distribution, yet remain relatively well regularized, thereby avoiding "over-fitting" and severe underperformance on testing data. Yet, critically, when relying on explicit feature selection, across all input space dimension sizes for the FFDM and MRI data, and when fewer than 9 features were selected for the US data, the MCMC-BANN's success was contingent upon the use of LSW-FS over ARD-FS. The MCMC-BANN severely underperformed when coupled with the ARD-FS, especially when limited to picking only a few features. The smaller $\mathrm{AUC}_{0.632+}$. and higher bootstrap variability (most dramatically evident for the lower input space dimensions), reveals limitations in ARD-FS ability to consistently identify smaller sub-sets of features capable of robustly contributing to the classification task. This limitation may be in part due to ARD's capacity for discovering non-linear associations, which may vary highly between different bootstrapped sub-samples, as well as its less direct approach (compared to LSW) in feature determination.

Turning to LDA, while not best suited to model the non-linear DR mappings, the robustness and stability of LDA shines when joined with LSW-FS for classification purposes. LDA is, in a sense, naturally regularized by its linear nature and thus automatically avoids severe over-fitting situations. Often, the relative advantage of a more complex classifier, such as MCMC-BANN, over LDA, may begin to erode as sample size decreases, even if the underlying distribution is not completely linear in nature. These phenomena are apparent for the much smaller FFDM (245 unique cases, on 735 images) and DCE-MRI (356 unique lesions/images) datasets, as the less sophisticated LDA often produced the highest $AUC_{0.632+}$ values. The LDA classifier showed the greatest strength with the MRI data, nearly matching the LSW-MCMC-BANN and similarly for the DR approaches.

Furthermore, in examining Figure 2 again, among points falling within desirable performance specifications (upper left-hand corner: high classification performance/lower expected variability), it is reasonable to favor configurations which require the lowest input space dimensionality, as discussed previously (either the number selected features or target embedded mapping dimensions). A potential advantage of DR is that it may reduce the amount of necessary parameters (not including the unsupervised transformation characterized by the data itself) required to form a satisfactory data representation suitable for robust classification. In fact, most motivation for performing DR is lost if the target dimension is not considerably lower than the original high dimensional space. This is because such mapped representations become less efficient compared to simply making use of the original feature space or selected sub-space as dimensions are added. Thus, within the framework of these criteria, in reviewing the results from the three modalities on whole, one may postulate, that as an overall strategy, 4D t-SNE appears likely to produce competitive classification performance when used as input into a non-linear classifier such as the MCMC-BANN. Such classification performance coupled with the intriguing 2D and 3D visualizations of the overall data structure may evoke attractive research potential.

In practice, it should be noted that, with the sole intention of maximizing classification performance based on finite sample training data, there may be no clear advantage for use of DR techniques over traditional feature selection. Although, again, due to the "curse of dimensionality," as the input space dimension for classification becomes higher in dimension, eventually cross-validation based-performance will stagnant or even begin to regress lower. This occurs as the dataset sample size is not sufficient to adequately isolate a unique classifier solution (as many, potentially infinite, become possible) and marginal, if not none at all, new information is gained by the additional dimensions. Thus, for these reasons and in order to compare each dataset on common ground, the tests were limited to 1D-13D.

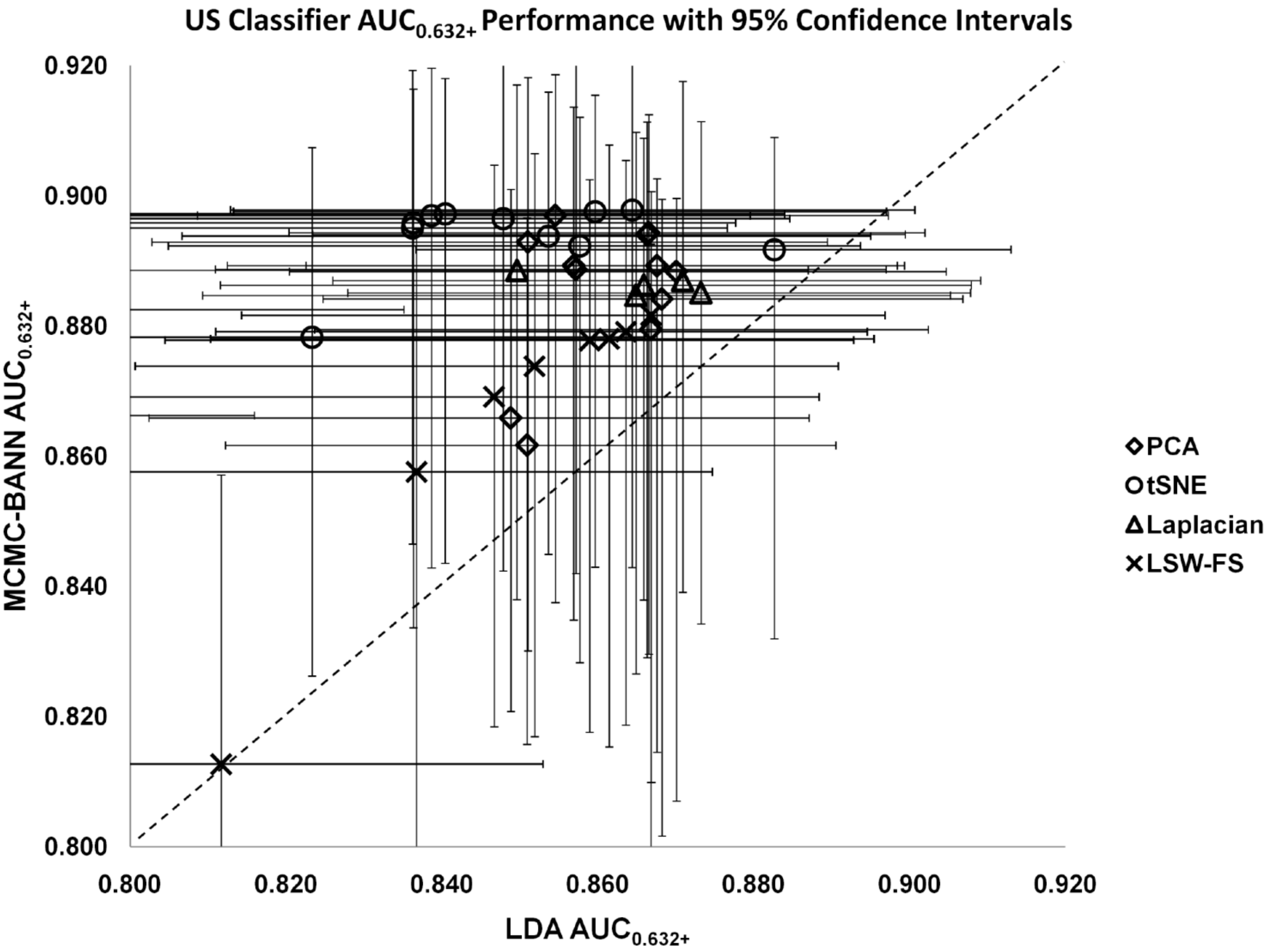


Figure 5. The 0.632+ bootstrapped area under the ROC curve is shown for MCMC-BANN (vertical axis) versus LDA (horizontal axis) with 95% empirical bootstrap confidence intervals included, for the originally 81 dimensional US feature dataset dimension reduced input or with LSW selected features.

VI. Conclusion

The ability to capture high-dimensional data structure in a human interpretable low-dimensional representation is a powerful research tool. The above findings strongly suggest the relevance of non-linear DR and representation techniques to future CADx research. DR cannot be expected to replace the benefits of feature selection based approaches in many cases. Yet, these techniques, in addition to competitive classification performance, do offer complementary information and a fresh perspective on interpreting the overall structure of the feature data. Of interest to future studies is to further investigate the origin, meaning, and physical interpretation of the discovered structures present in the CADx lesion data as revealed by these non-linear, local-geometry preserving representations. Such rich data structure representations may offer novel insights and useful understandings of clinical CADx image data.

VII. Acknowledgments

This work is partially supported by US DoD Grant No. W81XWH-08-1-0731, from the US Army Medical Research and Material Command, NIH grants P50-CA125138, and DOE grant DE-FG02-08ER6478. The authors would like to gratefully acknowledge Lorenzo Pesce, Richard Zur, Jun Zhang, and Partha Niyogi for their thoughtful discussion and insightful suggestions. Additionally, we thank Weijie Chen contributing breast MRI feature data. The authors are grateful to Geoffrey Hinton and Laurens van der Maaten for freely distributing their algorithm code as well as the very handy dimension reduction MATLAB toolbox. We would also like to gratefully acknowledge the SIRAF shared computing resource, supported in parts by grants NIH S10 RR021039 and P30 CA14599, and its excellent administrator, Chun-Wai Chan. And lastly we thank the reviewers for their useful suggestions.

VIII. Appendix

VIII. A. Laplacian Eigenmaps Algorithm Outline

Beginning with k input points, $x_1, \ldots, x_k$, in R^l:

Step 1: *Construct the Adjacency Graph:* Generate a graph with edges connecting nodes i and j if $\mathbf{x}_i$ and $\mathbf{x}_j$ are "close." Closeness is defined by the nodes included in the N nearest neighbors. This relation is naturally symmetric between points i and j. The parameter N must be selected.

Step 2: *Choosing Weight:* The "heat kernel" is used to assign weights to edge connected nodes i and j : $W_{ij}= exp(-||x_i-x_j||^2/t)$. Otherwise use $W_{ij} = 0$ for unconnected vertices. See Belkin and Niyogi for kernel justification[1]. The parameter t is user defined. If t is set very high, or approximately, $t = \infty$, the edge connected node weights are essentially $W_{ij} =1$, this option can be used to avoid parameter selection.

Step 3: *Computing Eigenmaps:* Assuming a connected graph generated in step 1, G, solve for the following eigenvector and eigenvalues: $Lf = \lambda Df$. , where D is the diagonal weight matrix, defined by summing over the rows of W . $D_{ii} =\Sigma_j W_{ij}$, and L is the Laplacian matrix defined as: $L=D-W$. Symmetric and positive semi-definite, conceptually the Laplacian matrix acts as an operator on functions defined by graph G's vertices. Solving the equation, let $\mathbf{f}_0,\ldots, \mathbf{f}_{k-1}$be the eigenvectors, arranged in accordance to their eigenvalues:$0 = \lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_k$. $L\mathbf{f}_0 = \lambda_0 D\mathbf{f}_0 \ldots L\mathbf{f}_{k-1} = \lambda_{k-1} D\mathbf{f}_{k-1}$.

Finally, the k input data points in R^l are embedded in m-dimensional Euclidean space using the m eigenvectors after the zero eigen-valued $\mathbf{f}_0$, $x_i \rightarrow (f_1(i),\ldots,f_m(i))$.

VIII. B. t-SNE Algorithm Outline

Beginning with k input points, $\{x_1, \ldots, x_k\}$ in R^l, set Perplexity parameter, *Perp,*

```
number of iterations $T$, learning rate $\eta$, and momentum $\alpha(t)$.
        Step 1. Compute Similarities: Compute pairwise $p_{j|i}$ probabilities using the $\sigma_i$ found with
                perplexity Perp, and use symmetrized conditional probability distributions
                $p_{ij} = (p_{j|i} + p_{i|j})/2k$
        Step 2. Initialize Solution Sample: Sample from $N(0,10^{-4}I^m)$ for initial points $\{y_1, ..., y_k\}$
        Step 3. Execute T Update Iterations on Y: Compute low-dimension similarities $q_{ij}$ using eq. (4)
                and gradient using eq(5).  Update Y using $Y^{(t)} = Y^{(t-1)} + \eta \frac{\delta C}{\delta y_i} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)})$
        Output: Low-dimension mapping $\{y_1, ..., y_k\}$ in $R^m$
```

IX. References

[1] M. Belkin, and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," Neural Comput. **15**, 1373--1396 (2002).

[2] L. van der Maaten, and G. Hinton, "Visualizing Data Using T-SNE," J. Mach. Learn. Res. **9**, 2605, 2579 (2008).

[3] M.L. Giger, H. Chan, and J. Boone, "Anniversary Paper: History and Status of CAD and Quantitative Image Analysis: The Role of Medical Physics and AAPM," Med. Phys. **35**, 5799-5820 (2008).

[4] Z. Huo, M.L. Giger, C.J. Vyborny, D.E. Wolverton, R.A. Schmidt, and K. Doi, "Automated Computerized Classification of Malignant and Benign Masses on Digitized Mammograms," Acad. Radiol. **5**, 155-168 (1998).

[5] M. Kupinski, and M. Giger, "Automated Seeded Lesion Segmentation on Digital Mammograms," IEEE Trans. Med. Imaging **17**, 510-517 (1998).

[6] Z. Huo, M.L. Giger, C.J. Vyborny, U. Bick, P. Lu, D.E. Wolverton, and R.A. Schmidt, "Analysis of Spiculation in the Computerized Classification of Mammographic Masses," Med. Phys. **22**, 1569-1579 (1995).

[7] K. Drukker, M.L. Giger, K. Horsch, M.A. Kupinski, C.J. Vyborny, and E.B. Mendelson, "Computerized Lesion Detection on Breast Ultrasound," Med. Phys. **29**, 1438-1446 (2002).

[8] W. Chen, M.L. Giger, U. Bick, and G.M. Newstead, "Automatic Identification and Classification of Characteristic Kinetic Curves of Breast Lesions on DCE-MRI," Med. Phys. **33**, 2878-2887 (2006).

[9] W. Chen, "Computerized Interpretation of Breast MRI: Investigation of Enhancement-Variance Dynamics," Med. Phys. **31**, 1076 (2004).

[10] K. Drukker, M.L. Giger, C.J. Vyborny, and E.B. Mendelson, "Computerized Detection and Classification of Cancer on Breast Ultrasound," Acad. Radiol. **11**, 526-535 (2004).

[11] M. Giger, "Computer-Aided Diagnosis of Breast Lesions in Medical Images," Comput. Sci. Eng. **2**, 39-45 (2000).

[12] G.D. Tourassi, B. Harrawood, S. Singh, J.Y. Lo, and C.E. Floyd, "Evaluation of Information-Theoretic Similarity Measures for Content-Based Retrieval and Detection of Masses in Mammograms," Med. Phys. **34**, 140-150 (2007).

[13] Y. Yuan, M.L. Giger, H. Li, K. Suzuki, and C. Sennett, "A Dual-Stage Method for Lesion Segmentation on Digital Mammograms," Med. Phys. **34**, 4180-4193 (2007).

[14] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic Press, Boston, 1990).

[15] C.M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006).

[16] S. Geman, E. Bienenstock, and R. Doursat, "Neural Networks and the Bias/Variance Dilemma," Neural Comput. **4**, 1-58 (1992).

[17] B. Sahiner, H. Chan, N. Petrick, R.F. Wagner, and L. Hadjiiski, "Feature Selection and Classifier Performance in Computer-Aided Diagnosis: The Effect of Finite Sample Size," Med. Phys. **27**, 1509-1522 (2000).

[18] M.A. Kupinski, and M.L. Giger, "Feature Selection with Limited Datasets," Med. Phys. **26**, 2176-2182 (1999).

[19] W. Chen, R.M. Zur, and M.L. Giger, "Joint feature selection and classification using a Bayesian neural

network with automatic relevance determination priors: potential use in CAD of medical imaging" in *Medical Imaging 2007: Computer-Aided Diagnosis,* edited by M. Giger and N. Karssemeijer (2007), vol.6514 of *Proc. SPIE, pp.* 65141G-10.

[20] M.A. Anastasio, H. Yoshida, R. Nagel, R.M. Nishikawa, and K. Doi, "A Genetic Algorithm-Based Method for Optimizing the Performance of a Computer-Aided Diagnosis Scheme for Detection of Clustered Microcalcifications in Mammograms," Med. Phys. **25**, 1613-1620 (1998).

[21] G.D. Tourassi, E.D. Frederick, M.K. Markey, and J. Floyd, "Application of the Mutual Information Criterion for Feature Selection in Computer-Aided Diagnosis," Med. Phys. **28**, 2394-2402 (2001).

[22] Y. Wang, D.J. Miller, and R. Clarke, "Approaches to Working in High-Dimensional Data Spaces: Gene Expression Microarrays," Br. J. Cancer **98**, 1023-1028 (2008).

[23] H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," J. Educ. Psychol. **24**, 498-520 (1933).

[24] M. Kirby, *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns* (John Wiley & Sons, Inc., New York, 2000).

[25] K. Drukker, N.P. Gruszauskas, and M.L. Giger, "Principal component analysis, classifier complexity, and robustness of sonographic breast lesion classification",*Medical Imaging 2009: Computer-Aided Diagnosis*, edited by M.Giger and N. Karssemeijer (2009), vol. 7260, *Proc. in SPIE*, pp. 72602B-6.

[26] C. Varini, A. Degenhard, and T.W. Nattkemper, "Visual Exploratory Analysis of DCE-MRI Data in Breast Cancer by Dimensional Data Reduction: A Comparative Study," Biomed. Signal Process. Control **1**, 56-63 (2006).

[27] A. Madabhushi, P. Yang, M. Rosen, and S. Weinstein, "Distinguishing Lesions from Posterior Acoustic Shadowing in Breast Ultrasound Via Non-Linear Dimensionality Reduction," Conf. Proc. IEEE Eng. Med. Biol. Soc. **1**, 3070-3073 (2006).

[28] M.K. Markey, J.Y. Lo, G.D. Tourassi, and C.E. Floyd, "Self-Organizing Map for Cluster Analysis of a Breast Cancer Database," Artif. Intell. Med. **27**, 113-127 (2003).

[29] K. Drukker, K. Horsch, and M.L. Giger, "Multimodality Computerized Diagnosis of Breast Lesions Using Mammography and Sonography," Acad. Radiol. **12**, 970-979 (2005).

[30] H. Chan, D. Wei, M.A. Helvie, B. Sahiner, D.D. Adler, M.M. Goodsitt, and N. Petrick, "Computer-Aided Classification of Mammographic Masses and Normal Tissue: Linear Discriminant Analysis in Texture Feature Space," Phys. Med. Biol. **40**, 857-876 (1995).

[31] B. Sahiner, H. Chan, and L. Hadjiiski, "Classifier Performance Prediction for Computer-Aided Diagnosis Using a Limited Dataset," Med. Phys. **35**, 1559 (2008).

[32] R.M. Neal, *Bayesian Learning for Neural Networks* (Springer-Verlag New York, Inc., 1996).

[33] M. Kupinski, D. Edwards, M. Giger, and C. Metz, "Ideal Observer Approximation Using Bayesian Classification Neural Networks," IEEE Trans. Med. Imaging **20**, 886-899 (2001).

[34] M.E. Tipping, in *Advanced Lectures on Machine Learning* (Springer , Berlin / Heidelberg, 2004), pp. 41-62.

[35] I. Nabney, *Netlab* (Springer, 2002).

[36] E. Levina, and B. Bickel, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2005).

[37] M. Belkin, and P. Niyogi, "Towards a Theoretical Foundation for Laplacian-Based Manifold Methods," J. Comput. Syst. Sci. **74**, 1289-1308 (2008).

[38] L. van der Maaten, "Matlab Toolbox for Dimensionality Reduction " (2008).

[39] G. Hinton, and S. Roweis, in *Advances in Neural Information Processing Systems 15* (The MIT Press, Cambridge, 2003), pp. 833-840.

[40] L. van der Maaten, "T-SNE Files" (2008).

[41] L.L. Pesce, and C.E. Metz, "Reliable and Computationally Efficient Maximum-Likelihood Estimation of "Proper" Binormal ROC Curves," Acad. Radiol. **14**, 814-829 (2007).

[42] C.E. Metz, "Basic Principles of ROC Analysis," Semin. Nucl. Med. **8**, 283-298 (1978).

[43] J.A. Hanley, and B.J. McNeil, "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," Radiology **143**, 29-36 (1982).

[44] B. Efron, and R. Tibshirani, "Improvements on Cross-Validation: The .632+ Bootstrap Method," J. Am. Stat. Assoc. **92**, 548-560 (1997).

END OF APPENDIX B

APPENDIX C
Manuscript in Publication: Medical Physics 37, 4155 (2010).

Title:

Enhancement of Breast CADx with Unlabeled Data

Authors: Andrew R. Jamieson, Maryellen L. Giger, Karen Drukker, and Lorenzo L. Pesce
Department of Radiology, University of Chicago, Chicago, Illinois 60637

Abstract:

Purpose: Unlabeled medical image data is abundant, yet the process of converting it into a labeled ("truth-known") database is time and resource expensive and fraught with ethical and logistics issues. We propose a dual-stage CADx scheme in which both labeled and unlabeled ("truth-known" and "truth-unknown") data are used. This study is an initial exploration of the potential for leveraging unlabeled data towards enhancing breast CADx.

Methods: From a labeled ultrasound image database consisting of 1126 lesions with an empirical cancer prevalence of 14%, 200 different randomly sampled sub-sets were selected and the truth status of a variable number of cases was masked to the algorithm, to mimic different types of labeled and unlabeled data-sources. The prevalence was fixed at 50% cancerous for the labeled data and 5% cancerous for the unlabeled. In the first stage of the dual-stage CADx scheme we term "transductive dimension reduction regularization" (TDR-R), both labeled and unlabeled images were characterized by extracted lesion features which were combined using dimension reduction (DR) techniques and mapped to a lower-dimensional representation. (The first stage ignored truth status therefore was an unsupervised algorithm.) In the second stage, the labeled data from the reduced dimension embedding was used to train a classifier towards estimating the probability of malignancy. For the first CADx stage, we investigated three DR approaches: Laplacian Eigenmaps, t-distributed stochastic neighbor embedding (t-SNE), and principal component analysis (PCA). For the TDR-R methods, the classifier in the second stage was a supervised (i.e., utilized truth) Bayesian Neural Net (BANN). The dual-stage CADx schemes were compared to a single-stage scheme based on manifold regularization (MR) in a semi-supervised setting via the LapSVM algorithm. Performance in terms of areas under the ROC curve (AUC) of the CADx schemes was evaluated in leave-one-out and .632+ bootstrap analyses on a by-lesion basis. Additionally, the trained algorithms were applied to an independent test data set consisting of 101 lesions with approximately 50% cancer prevalence. The difference in AUC (ΔAUC) between *with* and *without* the use of unlabeled data was computed.

Results: Statistically significant differences in the average AUC value (ΔAUC) were found in many instances between training with and without unlabeled data, based on the sample set distributions generated from this particular ultrasound dataset during cross-validation and using independent test set. For example, when using 100 labeled and 900 unlabeled cases and testing on the independent test set, the TDR-R methods produced average ΔAUC = 0.0361 with 95% intervals [0.0301; 0.0408] (p-value << 0.0001, adjusted for multiple comparisons, but considering the test set fixed) using t-SNE and average ΔAUC = .026 [0.0227, 0.0298] (adj. p-value << 0.0001) using Laplacian Eigenmaps, while the MR based *LapSVM* produced an average ΔAUC = .0381 [0.0351; 0.0405] (adj. p-value << 0.0001). We also found that schemes initially obtaining lower than average performance when using labeled data only, showed the most prominent increase in performance when unlabeled data were added in the first CADx stage, suggesting a regularization effect due to the injection of unlabeled data.

Conclusion: Our findings reveal evidence that incorporating unlabeled data information into the overall development of CADx methods may improve classifier performance by non-negligible amounts and warrants further investigation.

Keywords: semi-supervised learning, transductive learning, non-linear dimension reduction, computer-aided diagnosis, breast cancer, unlabeled data

I. Introduction

The rise of digital imaging followed by increased sophistication of image output and lowering cost of data storage has resulted in the accumulation of a substantial amount of clinical image information. This new reality provides ample opportunity for enhancing the development of computer-aided diagnosis (CADx) algorithms.[1] More robust methodologies can now be implemented due to the simultaneous increase in the size of training, testing, and validation image databases and the availability of images with higher information content. However, the algorithmic training of CADx is commonly implemented via supervised classification, which requires that "truth" (i.e., actual biological disease status such as "malignant" or "benign") be known for each image. Unfortunately, reliable "truth" labeling is seriously time and resource consuming and therefore acts as a limiting factor to databases sizes.[2] Even if the gathering of pathological, genetic and radiological information associated with each clinical case is expected to become more efficient, a relative abundance of readily available unlabeled (i.e., "truth-unknown" or probability of disease equal to prevalence), or incompletely labeled (i.e., "truth-partially-known" or probability of disease higher or lower than prevalence for each specific case), images is likely to persist in most research contexts. For example, in the clinic, patients may be referred to an imaging follow-up rather than a biopsy. From a practical standpoint it is wasteful to completely discard this information, as these images are likely to contain useful information as indicated, for example, by research suggesting that radiologists' decision making processes might be endlessly refined by exposure to both labeled (i.e. probability of disease equal to 0 or 1) and unlabeled image data, interpretable as a development of a general sense of familiarity with the structures contained in the image "space." [3]

Unlabeled image data can be regarded as a sample drawn from the underlying probability distribution marginalized over the combined class-categories, e.g., all cases ignoring whether they are "malignant" or "benign". A large and unbiased unlabeled database sample provides detailed knowledge of the inherent structure of the marginal distribution of the images, which can guide the subsequent design of supervised classification on labeled cases and perhaps improve performance. [4] In other words, the unlabeled data may help “regularize” the training of CADx algorithms. **Figure 1** illustrates these concepts.

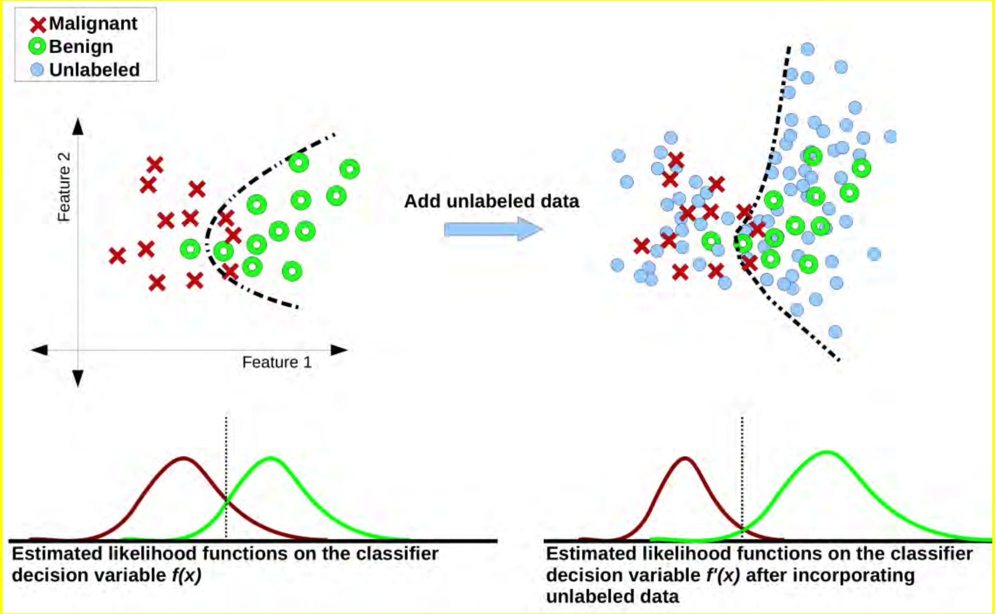

Malignant
Benign
Unlabeled
Feature 2
Feature 1
Add unlabeled data
Estimated likelihood functions on the classifier decision variable f(x)
Estimated likelihood functions on the classifier decision variable f′(x) after incorporating unlabeled data

Figure1.Simplified example illustrating how the use of unlabeled data might potentially improve CADx classifier regularization. The upper-left section displays a number of labeled samples from a hypothetical 2D feature space with a decision boundary (for likelihood ratio equal to 1) produced by a classifier trained on those data. The upper-right hand section depicts the same data, plus unlabeled samples which provide additional structural information, therefore altering the classifier and decision boundary. The lower section illustrates the class-conditional density functions of the classifier output decision variables obtained by applying the two trained classifiers as described above to the population.

The possibility for meaningful integration of unlabeled and labeled image data have been provided by "transductive" methods such as the recently developed unsupervised, local geometry preserving, non-linear dimension reduction (DR) and data representation techniques, including Laplacian Eigenmaps (Belkin and Niyogi) and t-distributed Stochastic Neighbor Embedding or t-SNE (van der Maaten and Hinton).[5-7] Additionally, building on the DR conceptual foundations for preserving inherent data structure, Manifold Regularization (MR) establishes the possibility for "truly" semi-supervised approaches, allowing for a natural extension to the immediate classification of out-of-sample test cases.[8] The purpose of our study is to introduce these methods to breast CADx and to provide a preliminary exploration of the potential for leveraging unlabeled databases towards the design of more robust breast mass lesion diagnosis algorithms. Additionally, the experimental design considered here aims to mimic, within the constraints imposed by the available dataset, clinically-relevant scenarios involving a potentially available unlabeled diagnostic datasets, specifically in terms of the expected cancer prevalence.

II. Background

II.A. Current Perspectives on Breast CADx

A detailed discussion of past and present breast image CADx methods can be found in a number of reviews.[1,9] A quick recapitulation suggests that these methods are intended to improve the quality and consistency of radiologists' clinical diagnoses and that they are usually designed following a supervised pattern recognition scheme constituted of segmentation, feature extraction, feature selection and classifier training/testing/validation. The relative merits of these steps are partially confounded by the limitation of utilizing relatively small datasets. Critical to the success of such methods are the informative value of the extracted features towards the specific diagnostic task, and the robustness of the classification algorithm employed to make use of the feature information. Feature selection (FS) is the final step of information evaluation and attempts to select the most discriminative input sub-space from a possibly large array of potential feature candidates. [10-12] An appealing alternative to explicit feature selection is to perform dimension reduction (DR), which we have previously compared with FS for multi-modality breast image CADx feature spaces including full-field digital mammography (FFDM), ultrasound, and dynamic contrast enhanced magnetic resonance imaging (DCE-MRI).[5] In this previous study, we evaluated classification performance and visualization of high-dimensional data-structures. The methods investigated, t-SNE and Laplacian Eigenmaps, are designed to discover the underlying structure of the data. Our analysis revealed that the DR methods, while not necessarily ready to completely replace FS, generally lead to classification performances on par with FS-based methods as well as providing 2D and 3D representations for aiding in the visualization of the original high-dimensional feature space.

II.B. Proposed Incorporation of Unlabeled Data for Training CADx

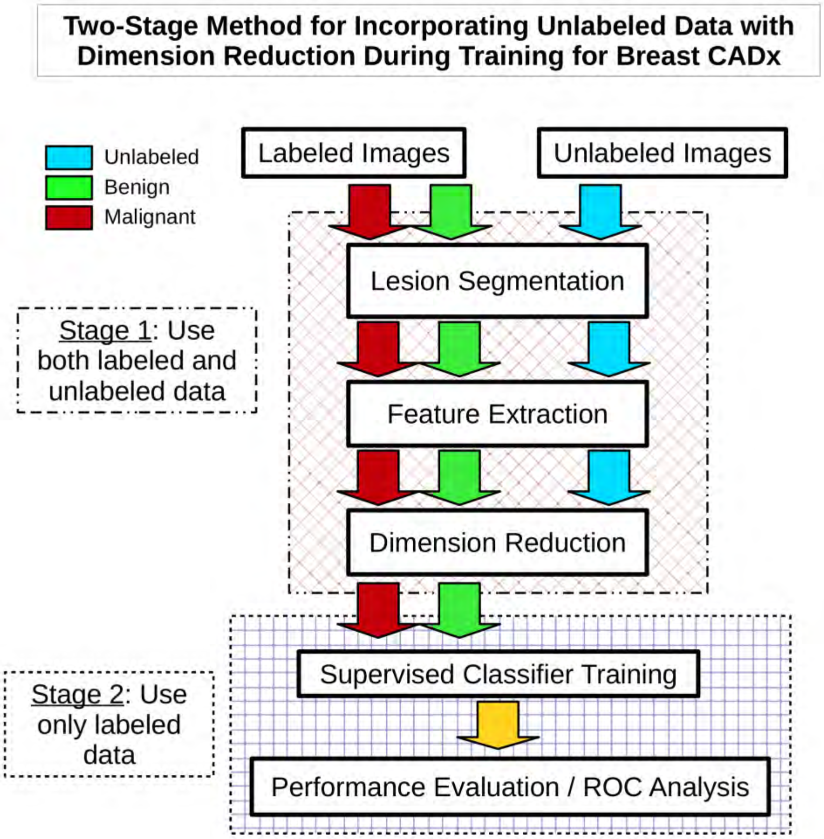

Two-Stage Method for Incorporating Unlabeled Data with Dimension Reduction During Training for Breast CADx
Unlabeled
Benign
Malignant
Labeled Images
Unlabeled Images
Lesion Segmentation
Stage 1: Use both labeled and unlabeled data
Feature Extraction
Dimension Reduction
Supervised Classifier Training
Stage 2: Use only labeled data
Performance Evaluation / ROC Analysis

Figure 2. Breast CADx algorithm work flow outline illustrating a two-stage method for incorporating unlabeled data with the use of dimension reduction.

In the previous work, we did not consider DR's capability of utilizing in a straightforward manner unlabeled data together with labeled data during the mapping from the higher to the lower-dimensional space. Since feature extraction is identical for labeled and unlabeled cases, instead of using supervised feature selection (such as automatic relevance determination), which is dependent exclusively upon the labeled cases, unsupervised dimension reduction can use the high-dimensional feature vectors, including the unlabeled feature data, to construct a lower-dimensional representation.[11] Ideally, the unlabeled data can help to more accurately capture the underlying manifold structure associated with the population of the imaged objects, even if some of the structure might not relate directly to the diagnostic task, e.g. describe differences among benign cases. **Figure 2** gives a broad outline of the proposed algorithm. We hypothesize that the labeled data sub-space produced by this type of DR mapping (including unlabeled data) could allow a supervised classifier to achieve enhanced classification performance. We call this approach "transductive-DR regularization" (TDR-R). The TDR-R approach requires the potentially computationally-intensive re-mapping step each time a new case is introduced. As differentiated from *supervised learning* which requires full knowledge of class categorization/labeling for training data, and *unsupervised* methods which do not use any information related to class identity, *semi-supervised learning* (SSL), in general, refers to a class of algorithms designed to make use of and learn from both labeled and unlabeled examples in a unified fashion for the task of classification.[4] We thus also included a “truly” semi-supervised learning algorithm known as Manifold Regularization (MR), which is designed to explicitly incorporate unlabeled data information data during training and can be extended to classify new cases without the re-mapping and re-training of transductive.[8]

II.C. Related Work Involving Unlabeled Data

To our knowledge, the use of non-linear, local geometry preserving DR and Manifold Regularization to exploit unlabeled image feature data towards improving breast lesion CADx classification performance has yet to be investigated. However, methods involving unlabeled data have been briefly investigated in the area of computer aided detection (CADe). Li and Zhou proposed to use unlabeled image data in conjunction with their algorithm "Co-Forest" — a modification of ensemble and co-training based learning techniques — for a CADe application focused on micro-calcifications in digital mammograms.[13] In their paper, the authors provided limited results based on an experimental design using only 88 images total. In the broader field of computer analysis in medical imaging, others have investigated the use of k-means clustering with texture analysis for unlabeled liver MRI image regions towards diagnosis of cirrhosis; unfortunately their conclusions were also limited because of their relatively small study size.[14] The use of unlabeled data information for classification tasks is a growing research interest outside of the medical imaging arena as well, for example in the analysis of protein sequences and speech/audio recognition.[15,16] Additionally, research exists on full image-space input based approaches (as opposed to using fixed pre-determined features), inspired in part by human-like visual systems that are intimately associated with the use of unlabeled stimuli.[17] Again, because of the relative abundance of unlabeled or incompletely labeled data in health-care related fields, such as image processing and CAD research, we expect that the challenge of how to effectively use such information will likely remain highly relevant.

III. Methods

III.A. Overview

Our experiments were based on sets of randomly selected cases from previously acquired retrospective datasets consisting of labeled cases. Each of the cases was represented by computer extracted features obtained from ultrasound (US) images of breast mass lesions. Each set consisted of labeled and "mock" unlabeled samples (i.e., cases for which the truth was ignored in that specific experiment for the purpose of assessing the effect of unlabeled data). For each experimental run, cases were selected, on a by-lesion basis, according to specific sampling criteria, including clinically relevant cancer prevalence percentages with respect to both the labeled and unlabeled data, as well as varying the total number of labeled and unlabeled cases used. After generating these samples, the algorithms were trained and tested, with and without the unlabeled data. The sub-sections below review our approach in detail.

III.B. CADx Breast Ultrasound Dataset

The ultrasound (US) data characterized in this study consists of clinical breast lesions presented in images acquired at the University of Chicago Medical Center. Lesions were labeled according to pathological truth — determined either by biopsy or radiologic report and collected under HIPAA-compliant IRB protocols.

The US image breast lesion feature datasets were generated from previously developed CADx algorithms at the University of Chicago. [18-21] Based on the manually identified lesion center, the CADx algorithm performed automated-seeded segmentation of the lesion margin followed by computerized feature extraction. Morphological, texture, and geometric features, as well as those related to posterior acoustic behavior were extracted from the images. Further details regarding the previously developed features used here can be found in the provided references.[18-20] **Table 1** summarizes the content of the US databases used, including the total number of lesion features extracted. The benign ultrasound lesions can be sub-categorized as benign solid masses and benign cystic masses. This study only considered the binary classification task of distinguishing between cancerous vs. non-cancerous (termed "malignant" and "benign") lesions. The empirical cancer prevalence for the first dataset was approximately 14% and 50% for the independent testing dataset. Again, all sampling and performance evaluations were conducted on a by-lesion basis, as multiple US images may be associated with a single unique lesion. In such an instance, classifier output from all associated images for a single physical lesion case is averaged.

Table 1. Feature Database Composition.

Dataset	Total Number of Images	Number of Malignant Lesions	Number of Benign Lesions	Total Number of Lesion Features Calculated
Training and Cross Validation Set	2956	158	968 (401 mass; 567 cystic)	81
Independent Test Set	369	54	47 (34 mass; 13 cystic)	81

Table.2 Summary of the four approaches explored for incorporating unlabeled data in breast CADx.

Method Type		Stage 1	Stage 2
		Unsupervised Dimension Reduction (DR)	Supervised Classifier
1	Transductive DR Regularization	PCA (linear)	BANN
2	Transductive DR Regularization	Laplacian Eigenmaps (non-linear)	BANN
3	Transductive DR Regularization	t-SNE (non-linear)	BANN
4	Manifold Regularization	Combined stages using Semi-Supervised algorithm: LapSVM	

III.C. Frameworks for Incorporating Unlabeled Data in CADx

General Framework

The approaches considered here build on the geometric intuitions motivating the design and use of non-linear DR techniques. This framework assumes that knowledge limited to the underlying marginal probability distribution, P_x, i.e., without labeling, can contribute towards identifying better classification decision functions for the task of modeling the conditional probability $P(y|x)$, where y is the target class label. This requires that if two points, x_1 and x_2 are close according to the intrinsic geometry of P_x, the conditional probabilities $P(y|x_1)$ and $P(y|x_2)$ are likely to be similar.[4] Algorithmic details applying this concept using two different techniques are provided below. It is important to note that all these methods assume that the unlabeled data are from the same underlying population as the labeled data and that both are unbiased samples (possibly conditional on truth for the labeled data). Therefore, in the form described here they are not designed to compensate for verification bias and similar sampling issues. Additionally, we note that for finite sample datasets, one cannot know with certainty if a sample satisfactorily represents the underlying population probability distribution. However, as the dataset size increases, the quality of the underlying marginal distribution representation is expected to improve.

i. Transductive DR Regularization (TDR-R) Approach

As previously stated, features are extracted in the same way for malignant and benign lesions as well as for and labeled and unlabeled lesions. Therefore unsupervised DR can be applied to datasets made of both labeled and unlabeled data in a straightforward manner, (*stage 1*, **Fig. 2, Fig. 3**). Next, a supervised classifier is trained using only the labeled samples, with feature information expressed in the reduced dimensionality representation (stage 2, **Fig. 2, Fig. 3**). Conceptually, the DR mapping acts as the agent through which the transductive learning principle is accomplished. Specifically, because the structure of the DR-generated "point-cloud" is dependent upon the presence of the unlabeled data, this influence acts as a regularizing force on the reduced-representation of the labeled cases, and hence the term "Transductive DR Regularization" (TDR-R). **Fig 3.** provides an overview of the training and testing (on an independent test dataset) of a breast CADx algorithm scheme incorporating TDR-R. It should also be noted that the TDR-R mappings considered here are in general non-parametric, reflected by the requirement they must be re-computed with each new set of data. In practice, a potential computational limitation may be incurred due to the requirement to re-compute the DR mapping for whenever new data needs to be analyzed. However, such concerns are expected to dissipate in time with the rapid, ongoing advance of computing technology, i.e. multi-core processors and "Grid" computing. Methods such as nearest neighbors approximations and the Nyström approximation can be used to estimate a lower dimensional mapping directly on new test data without including them into the DR process.[22] However, these approaches are not exact and often result in inconsistent performance. Thus, we decided to start exploring the potential of unlabeled data using transductive means. Because the test data must be introduced (albeit indirectly in an unsupervised fashion) during the training process, this approach is non-ideal and computationally costly. New approaches are under development aimed at overcoming these potential weaknesses.[23]

First Stage: Combining Labeled and Unlabeled Features in Unsupervised Dimension Reduction

Mathematically, the general problem of dimensionality reduction can be described as: provided an initial set $x_1, \ldots, x_k$ of k points in R^n , discover a set $x'_1, \ldots, x'_k$ in R^m , where $m << n$, such that x'_i sufficiently describes or "represents" the qualities of interest found in the original set x_i. For the specific context of high-dimensional breast lesion

CADx feature spaces, ideally, such lower-dimensional mappings should help to reveal relevant structural information associated with the categorization of the lesion sub-types and disease status for a population of breast image data.

Described briefly below are three DR techniques, one linear, and two non-linear, respectively: Principal component analysis (PCA), Laplacian Eigenmaps, and t-SNE. The latter two methods were chosen because of their distinct approaches to non-linearity and local structure. A brief description of these approaches is provided in the following, while a deeper discussion in the context of breast CADx can be found in our previous study.[5] Using this previous study as a heuristic reference point, in these experiments, beginning with all 81 features as initial input, the output reduced dimension was set to 7D for PCA, 5D for t-SNE and 7D for the Laplacian Eigenmaps. For the PCA and Laplacian Eigenmaps, we simply use the first consecutive output embeddings up to the dimension desired. For the t-SNE the output dimension is predetermined and all outputs are used. Details are described below.

PCA linearly transforms the input matrix of data into a new orthogonal basis set ordered according the fraction of global variance captured, in other words it performs an eigenvalue decomposition of the data covariance matrix.[24] Lacking the ability to explicitly account for non-linear and local structure, and hence assumed less likely to make efficient use of unlabeled data for regularizing labeled input used to train supervised classifiers, this linear dimension reduction method is included experimentally for comparison purposes only.

Building off of spectral graph theory, Laplacian Eigenmaps, proposed by Belkin and Niyogi, utilize the optimal embedding properties of the Laplace-Beltrami operator on smooth manifolds and its theoretical connections to the graph Laplacian.[25,26] Specifically, after a weighted neighborhood adjacency graph is formed using the original high-dimensional data space, eigenvalues and eigenvectors are computed for the graph Laplacian. Acting as a discrete approximation to the Laplace-Beltrami operator, the Laplacian of the point-cloud graph can be shown to preserve local neighborhood information optimally for some criteria, [25,26] hence motivating the use of its eigenfunctions in embedding into lower-dimensional spaces. [6] Two parameters are required to be set for Laplacian Eigenmaps: first, the number of nearest neighbors (NN) for constructing the connected graph, and second, the exponential heat kernel parameter, σ_{heat}. Based on our previous study [5], we chose NN=55 and σ_{heat} =1. Currently, no theoretical basis exists for univocal parameter selection.

The third method considered is t-Distributed Stochastic Neighbor Embedding (t-SNE), proposed by van der Maaten and Hinton. [7] Unlike the more theoretically motivated Laplacian Eigenmaps, t-SNE attacks DR from a probabilistic framework. The basis of t-SNE is to carefully define and compute pair-wise similarities between all points in the original high-dimensional space and then attempt to match this distribution in some lower-dimensional embedding by calculating a corresponding set of pair-wise similarities. The algorithm begins by randomly initializing points according to a Gaussian distribution in the lower-dimensional space, and then iteratively updates point positions by way of a cost function and update gradient based on the Kullback-Leibler divergence. Although such iterative and statistically-oriented approaches may require orders of magnitude more computational effort, greater flexibility and generality may be possible due to the easing of theoretical formalism, provided the system is well-conditioned. In the implementation used here, PCA is first applied to the data to accelerate convergence. In addition to the target embedding dimension, a single parameter called the *Perplexity* must be set which aids in the control of the local scaling used for the similarity calculations. This parameter was set to 30, following our previous paper. [5]

Second Stage: Using DR Mapped Labeled Cases in the Training of a Supervised Classifier

In order to perform supervised classification on labeled cases in the reduced mappings as noted in **Fig.2**, we implemented a Markov Chain Monte Carlo (MCMC) Bayesian artificial neural network classifier (BANN) using Nabney's *Netlab* package for MATLAB. [27] Provided sufficient training sample sizes, a BANN can be shown to model the ideal observer and achieve optimal classification, given a data source. [28] The network architecture consisted of the input layer nodes, a connected hidden layer with one node more than the input layer, and a single output target as probability of malignancy.

ii. "Truly" Semi-Supervised Learning with Manifold Regularization Approach

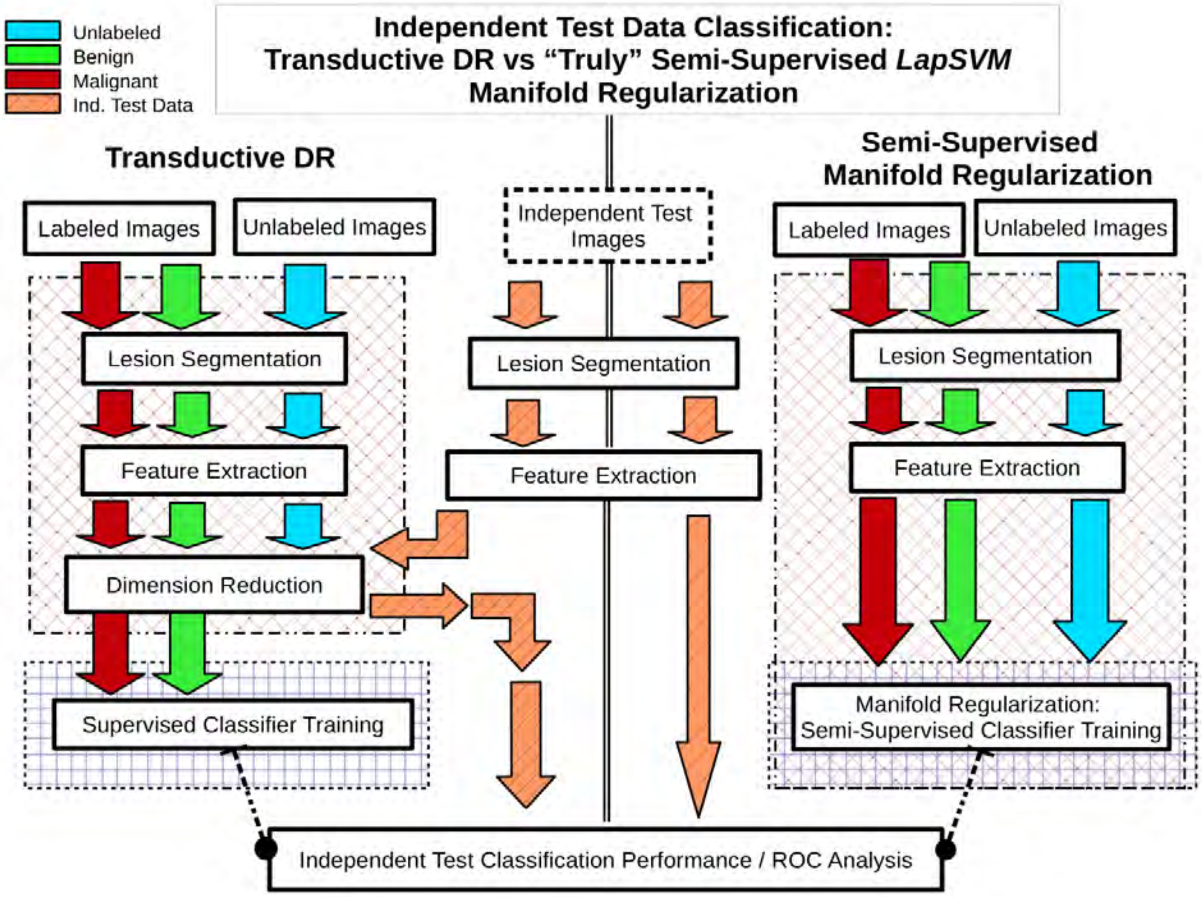


Figure 3. Schematic diagram illustrating the side by side comparison showing how new independent test data is handled for the TDR-R (left side) and MR (right side) algorithm work-flows that incorporate unlabeled data for breast CADx.

Belkin, Niyogi, and Sindhwani introduced the idea of Manifold Regularization (MR).[8] Using "Representer" Theorems and Reproducing Kernel Hilbert Spaces (RKHS), their key theoretical accomplishment was to discover functional solutions capable of both explicitly incorporating information from the intrinsic geometric structure of the data (including unlabeled data) and also naturally extending to the classification of future out-of-sample cases, without having to rely on transductive means. [8] **Fig. 3b**. illustrates a side by side comparison showing how new independent test data is handled by the respective TDR-R algorithm and the MR algorithm for CADx workflows. All 81 features extracted here are input into the MR algorithm. To briefly illuminate the nature of this latter approach, first we consider general supervised learning using only labeled data, which can be framed as the following problem:

$$f^* = \arg\min_{f \in H_K} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2. \qquad (1)$$

Equation (1) contains two terms, the empirical loss function (V), which attributes penalty cost for incorrect classification (e.g., the hinge loss: ($1 - y_i f(x_i)$)), and the regularization term $\|f\|_K^2$, which constrains the complexity of the function solution (f*), defined within the Hilbert space H_k. The relative penalty imposed on the "smoothness" of a function is controlled by the parameter γ_A. Notably, the penalty norm in eq.(1) is defined in what is called the *ambient* space, or the space in which the original data (in this case high-dimensional breast image CADx features) exist. Solutions of the form,

$$f^*(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x), \qquad (2)$$

where K is any positive semi-definite kernel can be found with the familiar convex optimization techniques used for RKHS based Support Vector Machines (SVM). [29]

Manifold Regularization works by including an additional term, $\gamma_I \|f\|_I^2$, which imposes a smoothness penalty on functions linked to the structure of the underlying lower-dimensional manifold geometry defined by the intrinsic structure of P_x.

$$f^* = \arg\min_{f \in H_K} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2. \qquad (3)$$

Depending upon whether the marginal distribution is known or unknown, Belkin, Niyogi, and Sindhwani provide a theoretical basis for expressing solutions in terms of RKHS-based functional forms. [8] Note that in the context of empirical sample-based applications, the true underlying distribution is not known, and thus an approximation to the intrinsic (i.e., properties are local and thus variable from point to point) geometry is required. Building off of the utility of Laplacian Eigenmaps for DR embedding, the intrinsic structure of the data is approximated with the graph Laplacian in a similar fashion as described above in Section **III.C.i**. This approximation is shown to also admit solutions in the familiar and convenient functional form of a RKHS, allowing for a relatively simple algorithmic implementation, as done in this research effort. The optimization problem for the approximate case is provided here:

$$f^* = \arg\min_{f \in H_K} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \frac{\gamma_I}{(u+l)^2} \mathbf{f}^T L \mathbf{f}, \qquad (4)$$

where $\boldsymbol{L}$ is the graph Laplacian, f is the decision function, and $1/(u+l)^2$ is the scaling factor for the Laplace operator. The u unlabeled samples are explicitly incorporated into the optimization problem above as well as in the associated solution, f^*, of form

$$f^*(x) = \sum_{i=1}^{l+u} \alpha_i K(x_i, x), \qquad (5)$$

where K is again any positive semi-definite kernel, and α the associated weighting coefficients. This solution can then be applied to classify independent test data.

“Truly” Semi-Supervised Learning (SSL) Classifier Algorithm: LapSVM

Since the solution to the above optimization problem admits the same form as standard kernel based approaches,[8] SVM algorithms can be extended to include intrinsic regularization, this is called *LapSVM*. Details of the algorithmic derivation can be found in the original publication. [8] We employed a MATLAB implementation of the *LapSVM* algorithm using radial-basis function (RBF) kernels and setting σ to 3. The graph Laplacian was built with Nearest Neighbors=25 and the heat kernel parameter set to 3.[6] Each time the *LapSVM* was trained, γ_A and γ_I in **eq.(4)** were adjusted according to the relative ratio of labeled and unlabeled cases. Note, when $\gamma_I = 0$, *LapSVM* reverts to the SVM solution. Although a vital component, it is important to note again that no theoretical formalism exists for optimal selection of the aforementioned parameters. We selected "reasonable" settings based off heuristic observations. Due to the finite sample size of the data, if attempts are made to tweak the parameter space excessively the risk of over-fitting may become significant. Because of this concern, we postpone a more thorough investigation of the parameter configurations to future simulation studies. Again, all 81 extracted features were input into the *LapSVM* algorithm.

Notably, the *LapSVM* can also be treated in a transductive fashion (similar to those schematics shown on the left in **Fig. 3**) by including the independent test set data into the graph Laplacian. This approach was investigated for comparison’s sake when testing the smaller ultrasound independent test data and will be referred to as *T-LapSVM.*

III.D. Experimental Design and Sampling Protocol

Different experimental configurations were considered in order to explore the possible impact of incorporating unlabeled ultrasound image feature data into CADx classification algorithms. Within the context of this classification task, we hypothesize that the two most important factors influencing performance are the number of cases involved and the prevalence of cancer for both the labeled and unlabeled datasets used, respectively. We attempted to mimic clinically-relevant situations to provide some guidance to the practical design and use of CADx systems.

Due to the finite size of the available ultrasound database used here, the scope of settings possible for our experimental design is restricted. Hence, beyond a point, scenarios involving a large number of labeled or unlabeled cases cannot be modeled reliably. Additionally, we were constrained by the inherent empirical cancer prevalence in our initial dataset. The cancer prevalence is approximately 14% for the entire 1126 case (2956 ROIs) diagnostic US feature data set **(Table 1)**. For the labeled supervised training/testing we focused on smaller set sizes of 50,100 and 150 lesions. Because the calculations are highly demanding, we explored only a limited number of unlabeled dataset sizes: small, medium, and as large as practically possible ($N_{UL} = 900$). The cancer prevalence was fixed at 50% malignant for the labeled case samples and 5% malignant for the unlabeled case samples (other prevalence configurations were considered, but were not included in this article due to length constraints). The table below summarizes the configurations considered.

	Number of Unlabeled Cases (UL)		
Number of Labeled Cases (L)	Small	Medium	Large

50 L	50 UL	500 UL	900 UL
100 L	100 UL	500 UL	900 UL
150 L	150 UL	400 UL	900 UL

Table 3. Summary of the experimental run configurations according to the number of cases used for labeled(L) and unlabeled(UL) datasets.

For each experimental configuration, 200 independently randomly sampled sub-sets were drawn, by-lesion, from the entire ultrasound feature dataset and identified to the algorithm as labeled or unlabeled according to the design specifications. For each sample set, the labeled and unlabeled sub-set of cases were forced to be mutually exclusive. Sampling was performed without replacement. It is important to accumulate an adequate number of samples to boost statistical power for identifying trends and overcoming the noise produced by inter-sample variability in performance due to the small dataset sizes, which is related to sampling distribution variability. Again, due to the finite dataset size limitation, it is important to note that for the larger unlabeled data sets (900 UL), the case composition will be highly similar between the larger sample sets. This is consistent with using the original dataset as the population because this limits feature values and their combinations in the sampled cases. On the other hand, this is a reasonably large dataset and the purpose of this paper was to explore the new methods with empirical data.

Lastly, we tested the effect of using unlabeled data during training on the separate independent test set (**Table 1**), obtained independently from the original larger dataset.

III.E. Performance Evaluation Methodology

The area under the Receiver Operating Curve ROC curve (AUC) was used to quantify classifier performance because it is not restricted to a specific and likely arbitrary operating point, sensitivity or specificity. Moreover, it usually provides larger statistical power. The AUC values were estimated using the non-parametric Wilcoxon statistic computed using libraries from the Metz's group at the University of Chicago.[30] Classification performance was estimated by leave-one-out (LOO), for the 50L and 100L experiments, and 0.632+ bootstrap (632+) cross validation (CV) for the 150L experiments, and the independent test set, all on a by-lesion basis. [31] For a given experimental configuration, for each of the 200 runs, the difference in the estimated AUC, ($\Delta AUC = AUC_{\text{with unlabeled}}-AUC_{\text{without unlabeled}}$), was found between classification performed with and without the use of unlabeled data. The paired, non-parametric Wilcoxon signed-rank test was applied to the ΔAUC values in each 200 run sets and to each of the sub-groups defined by the original AUC (without unlabeled) quartiles, i.e.: top 25th, top 25th to 50th, bottom 50th to 25th, and bottom 25th percentile. When necessary, p-values were adjusted for multiple comparisons testing using the *Holm-Sidak* step-down method.[32,33] Because of the considerable computational requirements, especially during cross-validation, the calculations were run on a local 256 CPU computing cluster. For example, while using an Intel Xeon E5472 CPU running at 3.0 GHz although the Laplacian Eigenmaps DR usually requires less than 15 seconds, the t-SNE DR can take over 15 minutes to complete on a 1000 case US dataset sample.

IV. Results

As an illustrative example, **Figure 4** displays the first three embedding dimensions produced (out of the 5D total) for the t-SNE DR mapping as well as AUC

$_{\text{LOO}}$ classification performance for a single data set run (out of the total 200 generated) with 100 labeled (L) cases, and 900 unlabeled (UL). The plot in **Fig. 4a** displays the t-SNE DR mapping produced with labeled data only, while for **Fig. 4b,c.** unlabeled data is incorporated during the mapping. For this particular single run, the estimated AUC $_{\text{LOO}}$ increased from 0.79 (SE=0.044) without the use of unlabeled data to 0.87 (SE=0.034) when unlabeled data is included during the DR mapping (these standard errors refer only to the test set variability, i.e, we are analyzing the performance of the trained classifiers and not the training protocol). Importantly, this run is a single positive performance change example and is not representative of the entire set of runs or average performance.

Estimated classification performance changes for the entire 200 runs and covers a wide range, as shown in scatter plots displayed in **Figure 5**. **Fig. 5** displays the AUC$_{0.632+}$ performance for 150L cases using 150 (blue), 400 (green), and 900 (red) UL cases for all 200 runs for each classifier. Both the cross validation (CV) and independent test (IT) results are shown for all methods, with the x-axis as the AUC *without* UL data, and the y-axis as the AUC *with* UL data. The thin diagonal line indicates equivalence between the two estimates.

A few observations can be made regarding these results. Overall, the t-SNE and Laplacian Eigenmap DR methods, **Fig. 5 c,d,e,f**, produced the largest variation (both positively and negatively) in AUC performance and, therefore, exhibited the noisiest distributions. For the cross-validation based performances, **Fig. 5 a,c,e**, i.e. except the *LapSVM*, it is difficult to discern which side of the diagonal the majority of points lie with an unaided eye. It is possible that the cross-validation procedure counter-acts or blurs out the changes produced by incorporating unlabeled data. Indeed, when using the independent test data, for all the methods except the linear PCA TDR-R, **Fig. 5b,d,f,h**, it is clearer that the majority of points reside on the upper side of the equivalence diagonal, indicating that the average AUC estimate obtained *with* unlabeled data is higher than that obtained *without* unlabeled data. For the independent test data, *LapSVM* most evidently displays an improvement in estimated AUC increase with the use of unlabeled data even if the estimated absolute AUC performance is reduced, which might be an indication that, for this specific instance, the *LapSVM* algorithm was more prone to over-fitting than the other three. Additionally, as indicated by the distinct layering of the blue, green, and red dots in **Fig. 5h,** it is clear that a higher amount of unlabeled data produces greater performance enhancement.

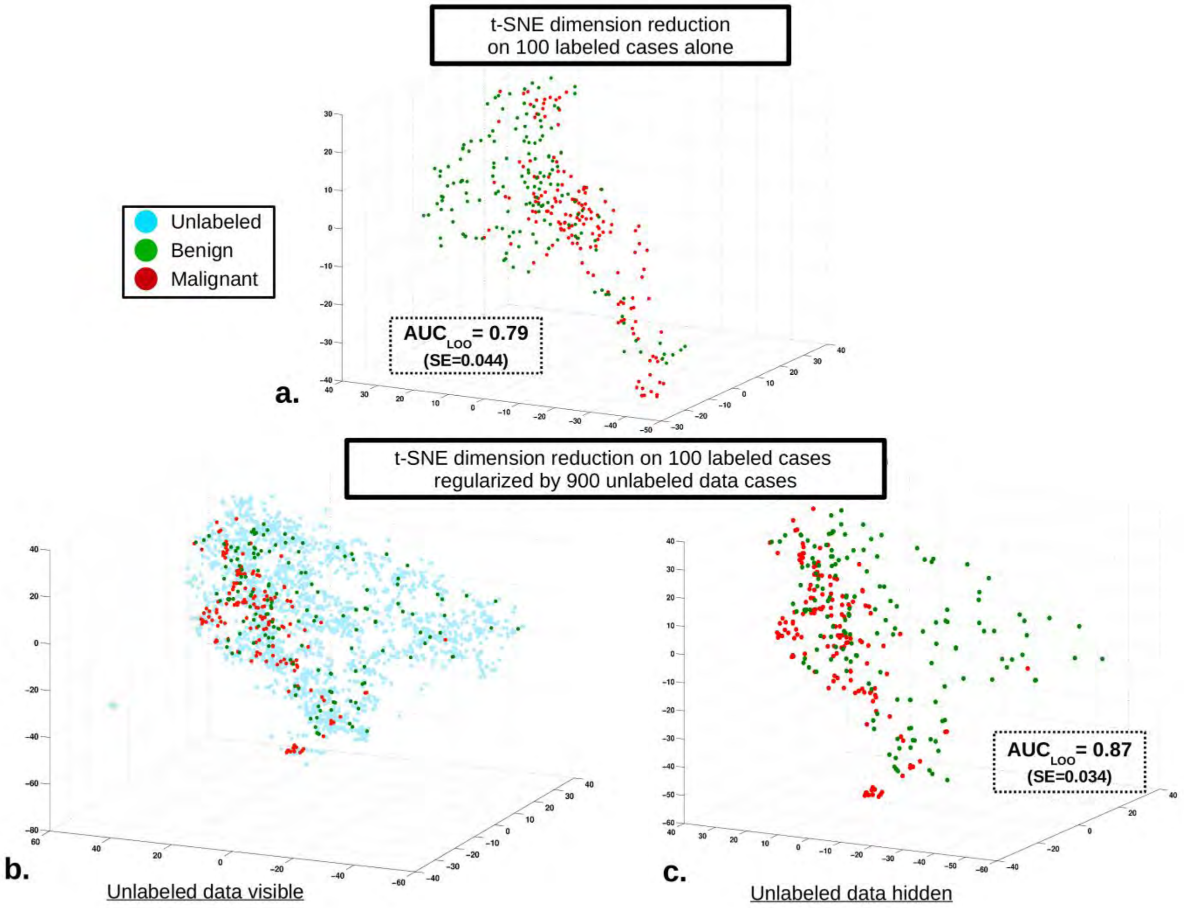


Figure 4. Example 3D visualization of the incorporation of unlabeled data for classifier regularization using t-SNE DR alongside the AUC $_{LOO}$ classification performance for a single run dataset (out of the total 200 generated) with 100 labeled (L) cases, and 900 unlabeled (UL). The three dimensions visualized are simply the first three embedding dimensions produced of the total 5D t-SNE DR. (a) Displays t-SNE DR mapping conducted with labeled data only, while for **(b,c)** unlabeled data is incorporated during the mapping. For this particular single run, classification performance as estimated by AUC $_{LOO}$ increased from 0.79 (SE=0.044) without the use of unlabeled data to 0.87 (SE=0.034) when unlabeled data is included during the DR mapping. However, this single run is not representative of the entire set or average performance, rather it is a single positive performance change example, a broad distribution of performances exists for the entire set of runs conducted, see **Fig.5**.

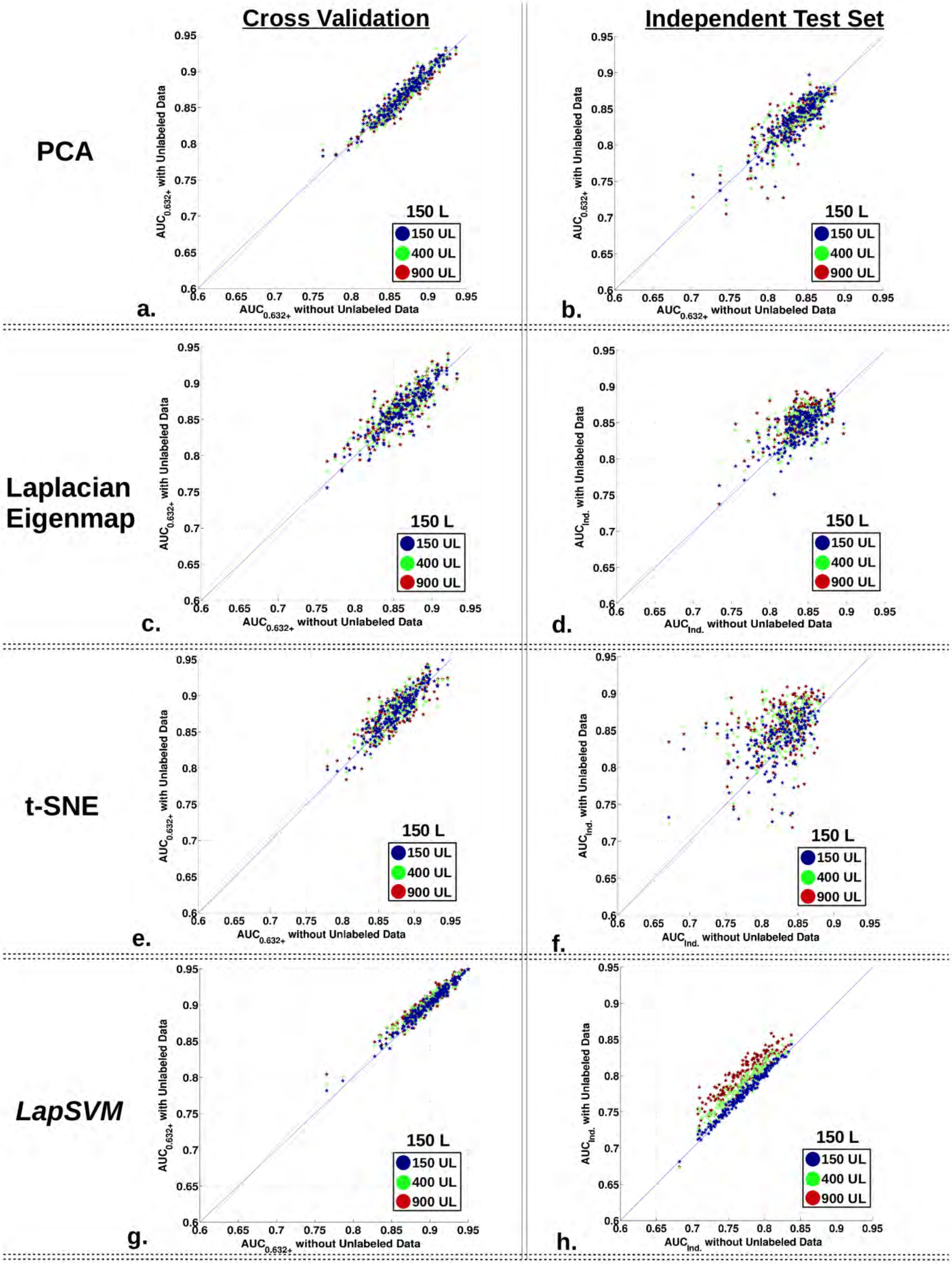

Cross Validation
Independent Test Set
PCA
Laplacian Eigenmap
t-SNE
LapSVM
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
$AUC_{0.632+}$ with Unlabeled Data
$AUC_{0.632+}$ without Unlabeled Data
$AUC_{Ind.}$ with Unlabeled Data
$AUC_{Ind.}$ without Unlabeled Data
150 L
150 UL
400 UL
900 UL
a.
b.
c.
d.
e.
f.
g.
h.

Figure 5. Scatter plots summarizing the classification performance distribution for the entire set of the 200 generated runs. The plots display the AUC$_{0.632+}$ performance for training with 150 labeled (L) cases using 150 (blue), 400 (green), and 900 (red) unlabeled (UL) cases for all 200 runs. The cross validation (CV) and independent test results are shown for all methods, (a,b) PCA, (c,d) Laplacian Eignemap, (e,f) t-SNE, and (g,h) *LapSVM* with the x-axis denoting the AUC without UL data, and the y-axis as the AUC with UL data for each run.

The AUC estimate distribution across the 200 generated runs can be condensed into a mean AUC and plotted according to the number of UL data included in the algorithm as shown in **Figure 6** for the use of 50L, 100L, and 150L cases across all classifier methods with associated error bars, based on the variance of the sample mean for the distribution of points, such as shown in the scatter plots (i.e. we are considering the large dataset as the population and ignoring validation-set variability because we are focusing on the effect this specific dataset). Additionally, statistically significant differences from ΔAUC = 0 for the average AUC are tabulated along with the rest of the results in **Table. 4,** including associated p-values adjusted for multiple-hypothesis testing by employing the *Holm-Sidak* correction**.** Consistent with the scatter plots in **Fig.5,** the influence of incorporating unlabeled data is most obvious for the independent set tests. For all the non-linear approaches, Laplacian Eigenmap, t-SNE, and *LapSVM*, the respective plots are positively sloped as more unlabeled data are added. Displaying this same trend most prominently, also included in **Fig.6,** are the results for transductive *LapSVM* or T-*LapSVM* on the independent test data. Notably, the linear PCA TDR-R appears relatively flat for both the cross validation and independent test set performance in **Fig 6**. Also, as see in **Fig. 6**, the mean AUC increases from approximately 0.78 at 50L, to 0.85 at 100L, and finally close to 0.90 for 150L for the LapSVM-CV. This trend clearly indicates the performance advantage of using more labeled data during training. For this dataset, on a per case basis, unlabeled data appears to have less impact on average performance gains. This is to be expected because unlabeled data lacks the variable that we are trying to predict: whether a case is cancerous or not. However, as mentioned earlier, unlabeled cases are frequently less resource consuming to acquire and put to use, and often a collection of unlabeled or poorly labeled data is readily available besides the labeled data.

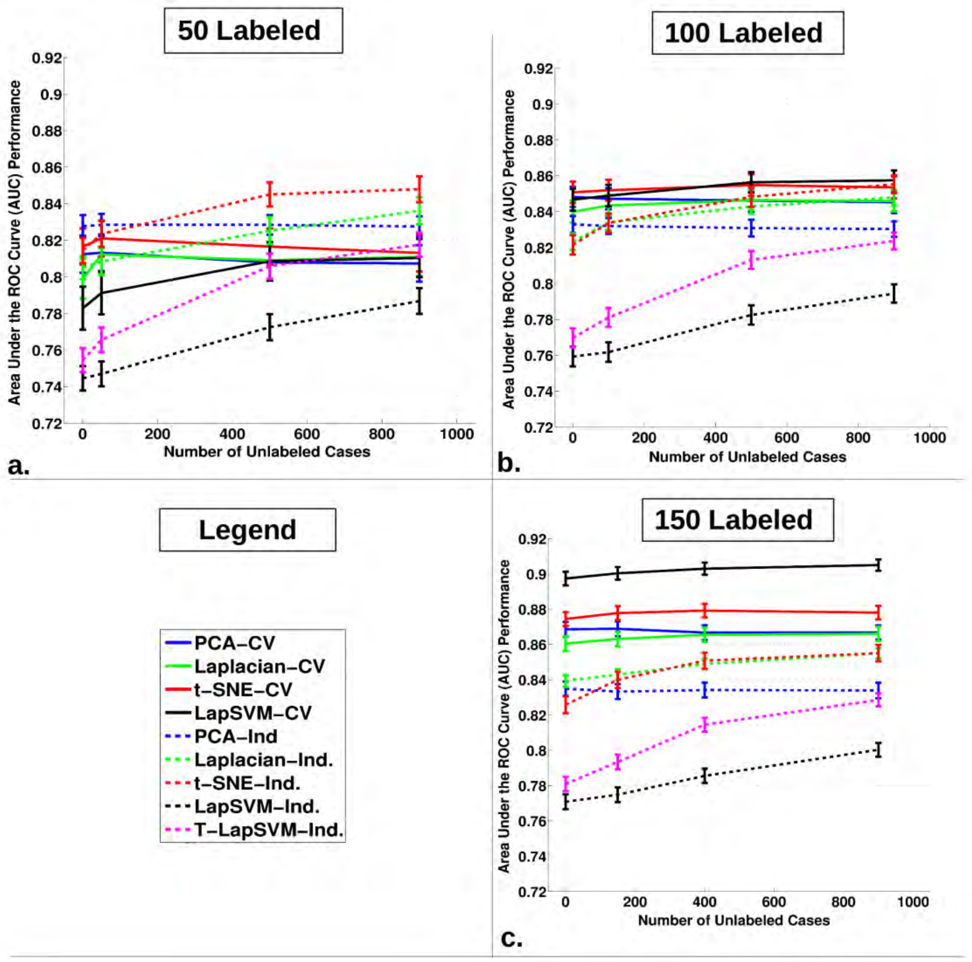


Figure 6. The average $\mathrm{AUC}_{\text{cross-validation}}$ and $\mathrm{AUC}_{\text{Independent.}}$ classification performance, with associated error bars, for all 200 runs, plotted against the number of unlabeled (UL) data incorporated in the given algorithm. Three plots are shown for (a) 50 labeled (L), (b) 100 L, and (c) 150 L cases including during the algorithm training respectively.

Cross Validation Results: Average ΔAUC

	Number of Cases			95% Conf. Int.			
Method	**Labeled**	**Unlabeled**	**Mean ΔAUC**	**Lower**	**Upper**	**Adj. p-value**	**Stat. Sig.**
TDR-R: PCA	50	50	-0.0007	-0.0033	0.0031	1	NO
	100	100	-0.0009	-0.0026	0.0019	1	NO
	150	150	0.0003	-0.0014	0.0012	1	NO
	50	500	-0.0057	-0.0086	-0.0009	1	NO
	100	500	-0.0020	-0.0038	0.0011	1	NO
	150	400	-0.0014	-0.0028	-0.0004	1	NO
	50	900	-0.0055	-0.0093	-0.0011	1	NO
	100	900	-0.0037	-0.0060	-0.0002	1	NO
	150	900	-0.0019	-0.0031	-0.0006	0.4733	NO
TDR-R: Laplacian	50	50	**0.0139**	0.0081	0.0196	0.0017	**YES**
	100	100	0.0035	0.0013	0.0069	0.8847	NO
	150	150	0.0026	0.0008	0.0047	0.8513	NO
	50	500	0.0088	0.0022	0.0143	1.0000	NO
	100	500	0.0062	0.0028	0.0098	0.1432	NO
	150	400	**0.0050**	0.0031	0.0074	0.0012	**YES**
	50	900	0.0122	0.0054	0.0175	0.0549	NO
	100	900	0.0060	0.0024	0.0097	0.2089	NO
	150	900	**0.0055**	0.0040	0.0084	0.0001	**YES**
TDR-R: t-SNE	50	50	0.0044	-0.0005	0.0091	1.0000	NO
	100	100	0.0017	-0.0020	0.0048	1.0000	NO
	150	150	0.0033	0.0017	0.0051	0.0492	**YES**
	50	500	0.0002	-0.0073	0.0064	1.0000	NO
	100	500	0.0061	0.0022	0.0102	0.4308	NO
	150	400	**0.0047**	0.0023	0.0068	0.0234	**YES**
	50	900	0.0001	-0.0066	0.0072	1.0000	NO
	100	900	0.0052	0.0010	0.0089	1.0000	NO
	150	900	0.0036	0.0012	0.0059	0.5995	NO
MR: LapSVM	50	50	**0.0084**	0.0067	0.0097	3.43E-18	**YES**
	100	100	**0.0022**	0.0018	0.0026	2.67E-17	**YES**
	150	150	**0.0030**	0.0022	0.0035	5.56E-11	**YES**
	50	500	**0.0259**	0.0208	0.0287	2.67E-21	**YES**
	100	500	**0.0105**	0.0088	0.0119	4.65E-22	**YES**
	150	400	**0.0056**	0.0046	0.0066	3.49E-17	**YES**
	50	900	**0.0287**	0.0222	0.0314	1.40E-20	**YES**
	100	900	**0.0117**	0.0094	0.0137	9.07E-17	**YES**
	150	900	**0.0070**	0.0057	0.0080	3.71E-19	**YES**

Table 4a. Results for the **average change in AUC** due to the use of unlabeled data are shown using **Cross Validation**. Included are the 95% confidence intervals and statistically significant differences from ΔAUC= 0 using a paired, non-parametric Wilcoxon signed-rank test, with consideration for multiple-hypothesis testing by employing the Holm-Sidak correction**.**

Independent Test Set Results: Average ΔAUC

	Number of Cases			95% Conf. Int.			
Method	**Labeled**	**Unlabeled**	**Mean ΔAUC**	**Lower**	**Upper**	**Adj. p-value**	**Stat. Sig.**
TDR-R: PCA	50	50	0.0015	-0.0012	0.0028	1.00E+00	NO
	100	100	-0.0007	-0.0032	0.0012	1.00E+00	NO
	150	150	-0.0015	-0.0037	0.0009	1.00E+00	NO
	50	500	0.0002	-0.0032	0.0027	1.00E+00	NO
	100	500	-0.0020	-0.0050	-0.0005	1.00E+00	NO
	150	400	-0.0009	-0.0027	0.0018	1.00E+00	NO
	50	900	-0.0014	-0.0055	0.0004	1.00E+00	NO
	100	900	-0.0026	-0.0057	-0.0006	1.00E+00	NO
	150	900	-0.0012	-0.0031	0.0018	1.00E+00	NO
TDR-R: Laplacian	50	50	**0.0046**	0.0029	0.0093	4.80E-02	**YES**
	100	100	**0.0119**	0.0089	0.0144	2.04E-10	**YES**
	150	150	**0.0048**	0.0027	0.0078	1.61E-02	**YES**
	50	500	**0.0235**	0.0207	0.0281	2.02E-19	**YES**
	100	500	**0.0207**	0.0180	0.0244	2.19E-18	**YES**
	150	400	**0.0108**	0.0073	0.0128	3.37E-09	**YES**
	50	900	**0.0333**	0.0310	0.0385	2.62E-23	**YES**
	100	900	**0.0260**	0.0227	0.0298	1.91E-19	**YES**
	150	900	**0.0169**	0.0140	0.0198	2.21E-17	**YES**
TDR-R: t-SNE	50	50	**0.0094**	0.0051	0.0131	2.16E-03	**YES**
	100	100	**0.0133**	0.0082	0.0165	6.83E-06	**YES**
	150	150	**0.0149**	0.0104	0.0187	9.79E-08	**YES**
	50	500	**0.0320**	0.0264	0.0361	8.78E-21	**YES**
	100	500	**0.0286**	0.0224	0.0330	3.05E-14	**YES**
	150	400	**0.0252**	0.0193	0.0283	1.34E-15	**YES**
	50	900	**0.0351**	0.0299	0.0389	2.29E-20	**YES**
	100	900	**0.0361**	0.0301	0.0408	3.00E-20	**YES**
	150	900	**0.0304**	0.0256	0.0345	1.43E-18	**YES**
MR: ***LapSVM***	50	50	**0.0026**	0.0017	0.0036	4.19E-05	**YES**
	100	100	**0.0033**	0.0023	0.0038	1.24E-10	**YES**
	150	150	**0.0050**	0.0041	0.0055	5.18E-24	**YES**
	50	500	**0.0309**	0.0281	0.0346	1.74E-27	**YES**
	100	500	**0.0252**	0.0224	0.0268	4.31E-29	**YES**
	150	400	**0.0177**	0.0160	0.0190	1.43E-30	**YES**
	50	900	**0.0467**	0.0428	0.0505	1.97E-31	**YES**
	100	900	**0.0381**	0.0351	0.0405	2.39E-30	**YES**
	150	900	**0.0334**	0.0311	0.0349	5.93E-32	**YES**

Table 4b. Results for the **average change in AUC** due to the use of unlabeled data are shown for the **Independent Test set** data. Included are the 95% confidence intervals and statistically significant differences from ΔAUC= 0 using a paired, non-parametric Wilcoxon signed-rank test, with consideration for multiple-hypothesis testing by employing the Holm-Sidak correction**.**

Only looking at the differences in average AUC ignores certain information-e.g., what is the effect of using unlabeled data on the variability of the resulting classifiers. As noted for **Fig.5,** due to the relatively small number of labeled cases used, a wide distribution of performances estimates is produced. Dividing the 200 run sets according to their initial performance quartiles (without UL data), as described previously, allows one to observe how the use of unlabeled data appears to affect the relatively under-average, average, or above-average performing classifiers each of which was trained with a given labeled dataset. The differential impact on performance caused by the incorporation of unlabeled data in these CADx schemes may consider classifier regularization effects in terms of whether differently performing classifiers tend to move closer to an average (and higher) performance after regularization. And while the restrictions of our finite datasets limit the generalizability of our results, we believe it is reasonable to assume that the overall trend in performance changes will reflect a more general property of this type of regularization.

Specifically, the initial AUC estimate performance distribution from the classifier *without* UL data was further decomposed in to respective quartiles: top 25th, top 25th to 50th, bottom 50th to 25th, and bottom 25th percentile. **Figure 7,8** displays the change in AUC (ΔAUC = $\text{AUC}_{\text{with unlabeled}}$ - $\text{AUC}_{\text{without unlabeled}}$) according to the quartile decomposition across all classifiers for both cross-validation and independent test sets. In each plot, the quartile dependent change in AUC is ordered according to the use of 50L, 100L and 150L data moving left to right. Within each sub-set group, the triplet represents the use of a low (50,100,150 UL), medium (400/500 UL), and high (900UL) number of unlabeled data. Statistically significant differences from ΔAUC = 0 using a paired, non-parametric Wilcoxon signed-rank test, with consideration for multiple-hypothesis testing by employing the *Holm-Sidak* correction, are indicated by the ***** above the bars in **Figs. 7,8,9**. (Tests are again based on the distribution of points, as described previously.)

The primary observation to be made from constructing the ΔAUC quartile decomposition is essentially that the use of unlabeled data most dramatically impacts the performance of the initially lower-than-average performing runs, suggestive of a potentially regularizing effect on the classifiers. As clearly indicated by the long dark blue bars in **Fig. 7; Fig. 8,** the incorporation of unlabeled data provided the strongest performance boost to runs originating in the lowest 25th quartile (blue bars). Furthermore, moving from the lower quartile to the upper quartiles, respectively, the relative influence caused by including unlabeled data on classifier AUC performance is weakened. Interestingly, for a limited group of experimental configurations, such as for t-SNE and Laplacian Eigenmap with 50L data shown in **Fig. 7c,e**, the upper quartiles actually appear to trend in the negative direction.

For the CV results **Fig.7 a,c,e**, it is apparent that the number of labeled data used to train impacts the consequent degree of change in AUC when UL data is added, with the largest differences appearing for when training with 50 L cases. However, with the independent data test, the effect of the number of labeled training cases was less pronounced, as seen in **Fig. 7 b,d,f.**, Turning to the impact of the number of unlabeled used, overall, especially for the independent test set such as for the *LapSVM* in **Fig. 8b.**, the magnitude of the ΔAUC trends upward as more unlabeled data is included.

Figure 8 displays the quartile decomposition in the comparison of all classifiers using 100 L cases during training and the highest number (900 UL cases) unlabeled data. Again, the independent data set reveal the largest changes in ΔAUC. **Fig.9** also supports the idea that the linear PCA TDR-R is relatively ineffective in making use of unlabeled when assessed using AUC values, showing no indication of a sizable regularization effect.

Lastly, it is important to again emphasize the nature of the results analyzed here and the interpretation of the statistical significance reported. As noted above, the difference in AUC between classifiers incorporating unlabeled data and those which do not, is based on 200 runs, each generated using samples from the same larger 1126 lesion US dataset. In the context of this experiment, the large dataset is regarded as the "population." Aside for the single independent test, experiments here do not (and could not) explicitly evaluate variability on expected performance changes by validation-sets. Thus, statistically significant differences discovered here may not necessary generalize to other US datasets at large.

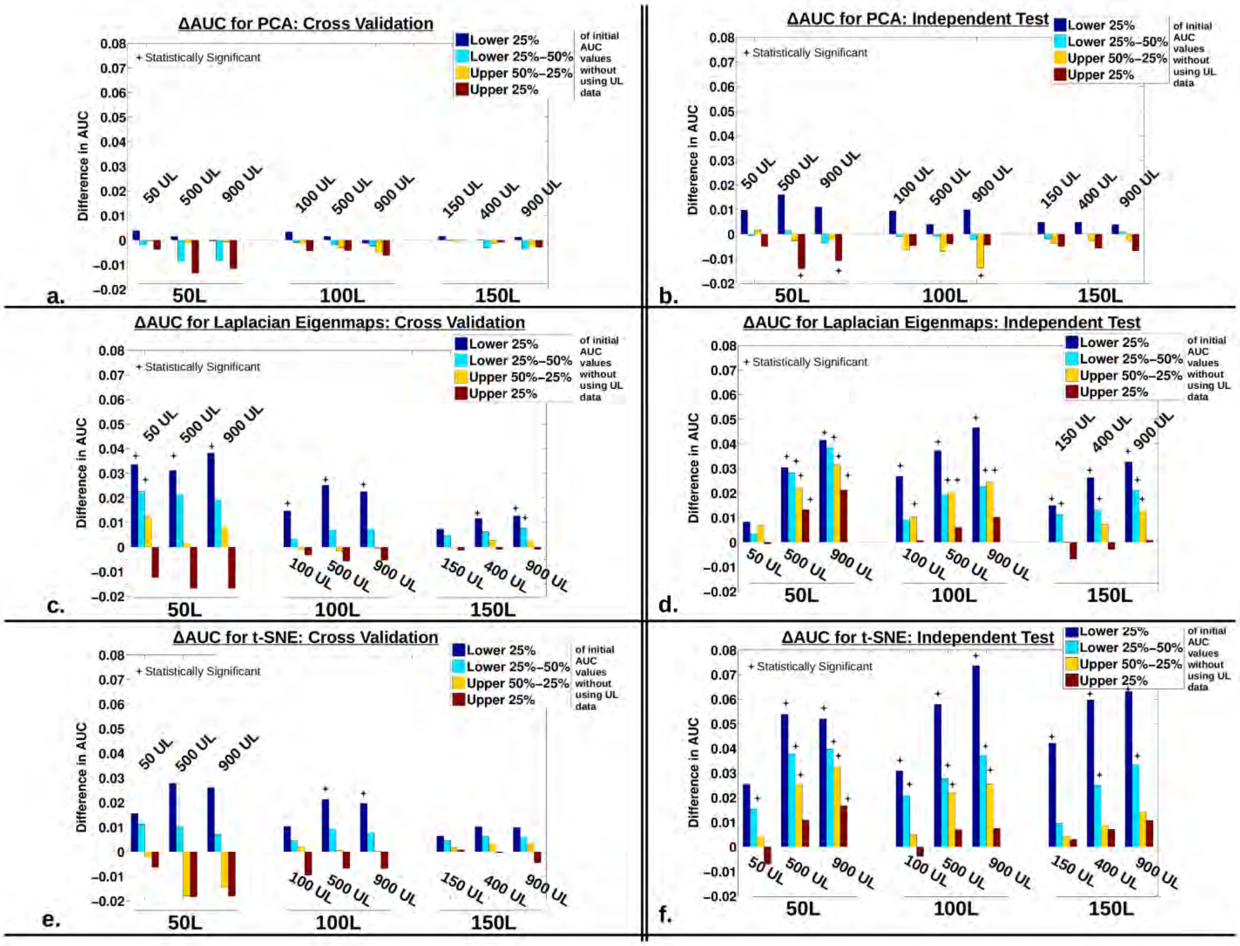


Figure 7. Results for the TDR-R methods, highlighting classifier regularization trends due to the use of unlabeled data. The difference in AUC (ΔAUC = AUC(with unlabeled(UL) data) – AUC (without UL data)) organized according to a quartile decomposition of the initial AUC performance without the use of unlabeled data (lower 25% in blue, lower 25%-50% in light blue, upper 50%-25% in orange, and upper 25% in dark red. In each plot, the quartile dependent change in AUC is ordered according to the use of 50L, 100L and 150L data moving left to right, during training. And within each sub-set group, the triplet

represents the use of a low (50,100,150UL), medium (400/500UL), and high (900UL) number of UL data. Statistically significant differences from ΔAUC= 0 using a paired, non-parametric Wilcoxon signed-rank test, with consideration for multiple-hypothesis testing by employing the *Holm-Sidak* correction, are indicated by the * above the bars (setting α = 0.05 or for adjusted p-values < 0.05). The plots are organized by the respective techniques, (a,b) PCA, (c,d) Laplacian Eignemap, (e,f) t-SNE, with cross-validation performance in the left column and the independent test set on the right.

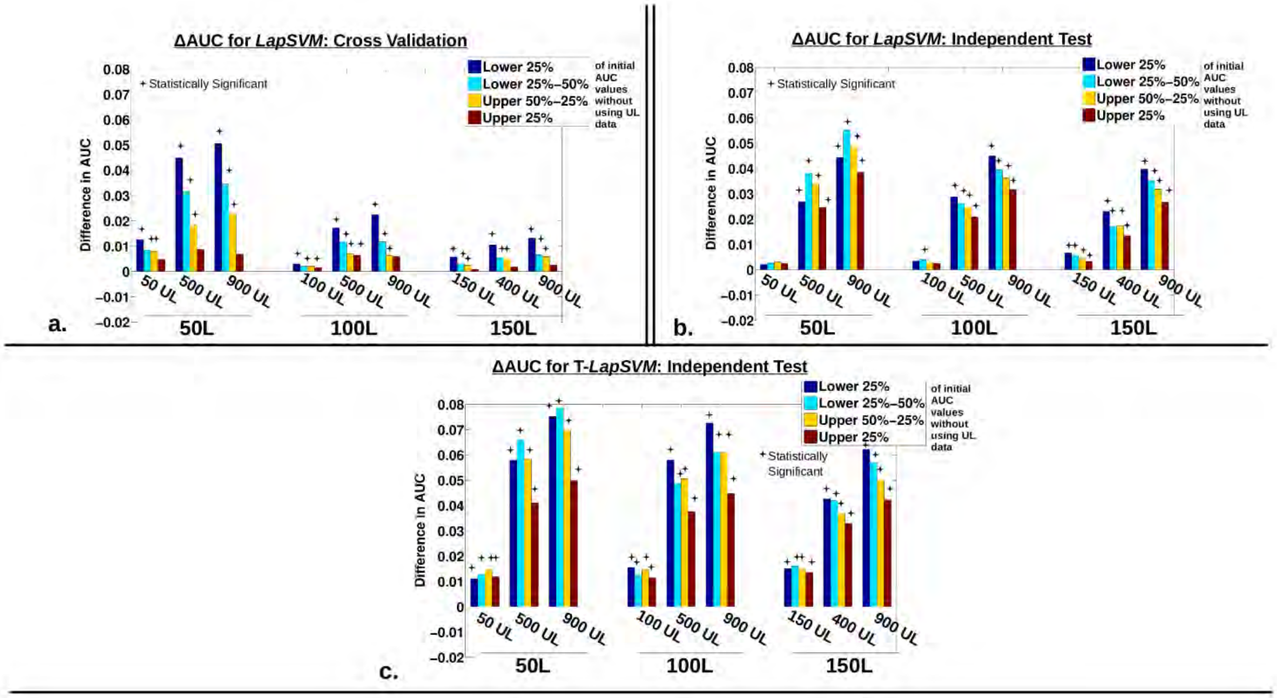


Figure 8. Results for the MR-based methods, highlighting classifier regularization trends due to the use of unlabeled data. The difference in AUC (ΔAUC= AUC (with unlabeled (UL) data) – AUC (without UL data)) organized according to a quartile decomposition of the initial AUC performance without the use of unlabeled data (lower 25% in blue, lower 25%-50% in light blue, upper 50%-25% in orange, and upper 25% in dark red. In each plot, the quartile dependent change in AUC is ordered according to the use of 50L, 100L and 150L data moving left to right, during training. And within each sub-set group, the triplet represents the use of a low (50,100,150UL), medium (400/500UL), and high (900UL) number of UL data. Statistically significant differences from ΔAUC= 0 using a paired, non-parametric Wilcoxon signed-rank test, with consideration for multiple-hypothesis testing by employing the *Holm-Sidak* correction, are indicated by the * above the bars (setting α = 0.05 or for adjusted p-values < 0.05). The plots display, (a), *LapSVM* for cross validation , (b) *LapSVM* with the independent test, and (c) T-*LapSVM* on the independent test.

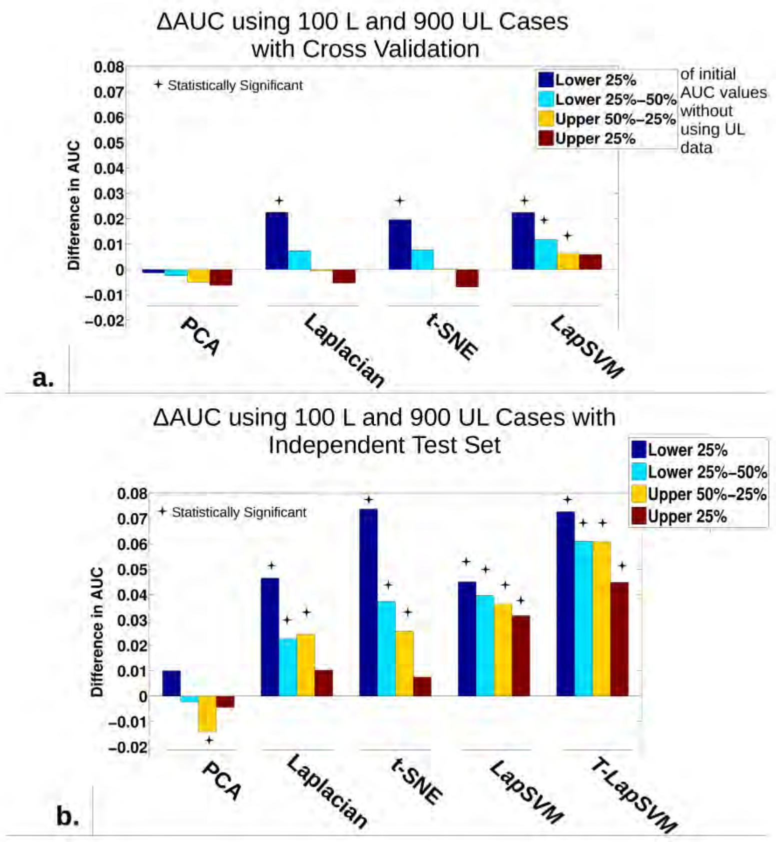


Figure 9. Using 100 L cases during training and highest number of unlabeled (UL) cases (900 UL), displayed are the differences in AUC (ΔAUC = AUC (with UL data) – AUC (without UL data)) organized according to a quartile decomposition of the initial $\text{AUC}_{\text{CV/Ind}}$ performance without the use of unlabeled data (lower 25% in blue, lower 25%-50% in light blue, upper 50%-25% in orange, and upper 25% in dark red, highlighting classifier regularization trends. Statistically significant differences from ΔAUC= 0 using a paired, non-parametric Wilcoxon signed-rank test, with consideration for multiple-hypothesis testing by employing the *Holm-Sidak* correction, are indicated by the * above the bars (setting $\alpha = 0.05$ or for adjusted p-values < 0.05). The plots show (a) the cross-validation performance, and (b) the independent test set performance.

V. Discussion

General Observations on the Use of Unlabeled Data

Overall, the above results provide evidence that classification performance is potentially enhanced by incorporating unlabeled feature data during the training of breast CADx algorithms. In particular, while the change in the mean AUC due to adding UL data appeared modest relative to the impact of using more labeled data, statistically significant results were found for both the cross-validation and the independent test set evaluations. Interestingly, after further analysis of the results via the quartile decomposition, a more detailed understanding of the nature of the performance changes was developed. Specifically, chief among the observations presented above, is that classifiers trained with a labeled sample set producing lower than average performance (using cross-validation or independent test data) were more likely to be positively impacted consequently by incorporating unlabeled data via either the TDR-R or MR based approaches. We interpreted this trend as a manifestation of the more general regularization properties one might expect to encounter by using unlabeled data in such a CADx scheme. We speculate that these observations may be consistent with the hypothesis that incorporating UL data via the use of structure-preserving DR techniques may help to more completely capture the inherent underlying distribution and thus render the classifiers trained on different samples more similar. Assuming such a theory to be true, the injection of UL data would most strongly impact sample sets which represent "poor" empirical estimations of the true underlying distribution and hence initially more likely to lead to trained classifiers with lower relative generalization performance. Consequently, the incorporation of the UL data, by aiding in more accurately capturing the inherent geometric structure of the data, could be construed as a beneficial regularizing influence on classifier performance. Conversely, for labeled sample sets which are more consistent with the inherent distribution, the introduction of additional UL cases would yield less enhancement, if any at all. Future investigations, and in particular simulations, are under way to answer these questions in more detail.

Performance Comparisons of the Different Approaches

The PCA TDR-R based approach appeared least capable of using unlabeled data. This result was expected as PCA is linear and cannot make efficient use of local and non-linear geometric qualities in the data structure, including when large amounts of UL data are present. Additionally, as suggested by the quartile decompositions, **Fig.7,8,9** PCA TDR-R did not appear to exhibit regularization trends present in the other methods. On the other hand, of the other approaches evaluated here, the MR *LapSVM* and T-*LapSVM* algorithms exhibited the most evident capacity for using unlabeled data to enhance classification performance. Specifically, as characterized in **Fig.5 g,h** , the classification performance of the *LapSVM* nearly always improved by incorporating unlabeled data. Furthermore, in addition to producing the "cleanest", least noisy scatter plots, the *LapSVM* showed the clearest differentiation in the relative performance enhancement for different amounts of UL data added, as seen in the layering of the blue, green, and red points on **Fig. 5h**. These results are perhaps not totally unexpected as the *LapSVM* algorithm is more sophisticated and theoretically grounded in its design for the explicit use of unlabeled data compared to the more heuristic TDR-R based approaches considered here. It should also be noted that when only using labeled data (that is 0 UL , e.g. the left most point on plots found in **Fig.6**) the *LapSVM* mimics a plain SVM classifier using all 81 features as input. Along these lines, as mentioned earlier we had previously shown that regularized classifiers using a large number of input features will perform similarly to classifiers trained on DR representations of the same features. [5]

However, while displaying a strong boost in estimated performance from the inclusion of unlabeled data, the *LapSVM* produced a lower absolute AUC performance on the independent test set compared to the other methods. It is not immediately clear why the *LapSVM* under-performed compared to the other methods with the independent test data. One possibility is that the kernel and Laplacian parameters used were not optimal for the independent data set. It is possible that the TDR-R methods partially avoided this dilemma by imposing stronger point-by-point regularization due to the requirement for generating a new reduced mapping when including the independent test data (which could also bias their performance evaluation making them look better on the independent test set because of that). In order to avoid further biasing the results and over fitting the algorithm to the data, we did not attempt to adjust or tweak any parameters during the performance evaluation on the independent test data set for any of the methods, and preserve the "one-shot" testing nature. This specific dilemma raises the more general and theoretically difficult problem of choosing appropriate parameters for techniques involved with manipulating and making use of unlabeled data or other unsupervised type tasks, such as clustering and DR. Moreover, this suggests that one should be very careful when assessing the performance of such an algorithm. These problems are active topics in machine learning research and we anticipate further advancements to be made in the near future. [34] Due to the current lack of adequate guidance on these issues, we identified this problem as beyond the scope of this manuscript. We are planning future simulation studies to more thoroughly investigate these theoretically oriented problems and how to possibly optimize the use of unlabeled data sets. We note that the primary objective of our effort here was to introduce these methods to breast CADx and provide a preliminary evaluation of the potential for using unlabeled data in the improvement of classification performance.

It should also be noted that in general there is no reason to assume an independent test sample should necessarily produce high performance, even when classified by the optimal Ideal Observer. This is because the independent test set may simply consist of samples from a poorly separating sub-space. In fact, as shown here in the independent test results, **Fig. 5 (**dotted lines**)**, as the labeled training set size is increased (50L, 100L, 150 L), although the variance decreases, the mean performance increased only slightly or not at all across all methods. This trend contrasts to the cross validation results (**Fig. 5** – solid lines), in which the mean estimated AUC classification performance continued to rise considerably as the training set size is increased. This is expected as cross validation methods, in addition to accounting for training and testing variability, attempt to estimate the expected performance of a classifier on the "population". Thus, as more training cases are available both variability and expected classification performance on the population should improve.

Impact of Cancer Prevalence

In our experience, the cancer prevalence in the labeled training set has a limited effect on classifier performance, unless the dataset is extremely unbalanced (very low or very high prevalence). The lower cancer prevalence in the unlabeled set reflects the fact that in clinical practice a 'hard' truth based on biopsy or surgery is much more likely to be unavailable for benign appearing lesions than for malignant looking ones. If lesions appear sufficiently benign, they are often assigned for imaging follow-up and not all of these will be processed to be included in a clinical database (too expensive and time consuming), while those that appear to be cancerous will be biopsied and included. Of the lesions assigned to imaging follow-up, a few may be missed cancers, while of the biopsied lesions, a certain fraction will turn out to be benign. Hence, there is a 'natural' division into how labeled (i.e., biopsied) cases and unlabeled cases are processed in clinical practice, which will produce a different prevalence (and might produce a bias if not done carefully). The majority frequently is unlabeled depending on the biopsy/recall rate of a given institution. Although only results for a single cancer prevalence (50% and 5% malignant, respectively, for the labeled and unlabeled sets) were shown here, other cancer prevalence settings were investigated. Further results were suppressed for this article as the presented findings were representative of the general trends, i.e., performance characteristics were not found to change in any considerable between the different cancer prevalence configurations. While this study did not reveal any overwhelming and immediately obvious trend associated with variation in cancer prevalence and the use of unlabeled data, as a general and unavoidable limitation to the overall study conducted here, the restriction of working with a finite data set available may have limited the statistical power required to clearly observe underlying differences due to cancer prevalence. Despite these initial findings, we believe that cancer prevalence and more generally the composition of categorical lesion sub-types and structure of the population space (such as ductal carcinoma in situ, cystic, infiltrating ductal carcinoma, etc.) which make up a set of feature data, may be of fundamental importance and potentially of critical interest to understanding how to use most effectively make use of unlabeled data in future work, including practical/clinical circumstances. Along these lines, it is of interest to consider how one might apply as additional input for training a potentially more robust classifier, the use of estimated prior, partial, or incomplete information (such as genetic, ethnic, risk characteristics) associated with an unlabeled data distribution when coupled to an existing known labeled data set. Additionally, it is worth investigating whether certain types of CADx data may be more amenable to the usage of unlabeled than others.

Clinical Relevance and Future Considerations

For the specific methods considered here, the MR *LapSVM* was currently the most practical candidate algorithm for clinical type situations, as it may be trained only once with inclusion of the unlabeled data and then later used to classify new independent test data without re-training. However, as the reality of affordable “desktop supercomputers” and scalable, real-time “grid”/”cloud” computing emerges, computational demands may be of less concern.[35] In fact, there may be definite advantages to conducting more computationally intensive, full transductive DR based approaches when analyzing new test data. The use of TDR-R based techniques, such as those employing t-SNE or Laplacian Eignemaps (or other DR-based methods not considered here), may offer useful visualization, such as for the example in **Fig. 4,** of the comparative structure and relative geometric orientation of newly acquired UL or new test cases added along with the original known data structure. It should also be noted that because the t-SNE and Laplacian Eignemaps approach the DR problem via distinct algorithmic mechanics, complementary information may also be gathered by combining both techniques in some fashion. As hinted in our previous article, such an evaluation may provide at least qualitative, but also, as techniques continue to mature, potentially quantitative, insight into the nature of the new data sets. [5] One such step in this direction is the recent proposal for a parametric t-SNE DR using deep neural networks. [23]

Lastly, we wish to emphasize again an important point. For most realistic scenarios, labeled data will almost always be more effective at improving performance than the same amount of unlabeled data. However, even if the “per case” utility of unlabeled data is only a fraction of that for labeled data, we believe the abundance of unlabeled available data, due to modern radiology practice, will provide sufficient impetus, in many contexts, to motivate exploitation of such nascent information.

VI. Conclusions

In summary, the incorporation of unlabeled feature data for the purpose of enhancing classification performance in the context of breast CADx was explored on four different algorithms. As discussed above, the results provide support for the hypothesis that including unlabeled data information during classifier training can act as a regularizing influence over cancer classification performance. The main limitation of this current study was the restriction of a finite, albeit relatively large, clinical database. However, we believe our results motivate future studies, both with simulations and using larger real clinical data sets. We expect a growing focus on such methods in the CADx research community with time.

VII. Acknowledgments

This work is partially supported by US DoD Grant No. W81XWH-08-1-0731 from the US Army Medical Research and Material Command, NIH grants P50-CA125138, and DOE grant DE-FG02-08ER6478. We would also like to gratefully acknowledge the University of Chicago SIRAF shared computing resource, supported in parts by grants NIH S10 RR021039 and P30 CA14599, and its excellent administrator, Chun-Wai Chan. The authors are grateful to G. Hinton and L. van der Maaten, as well as M. Belkin, V. Sindhwani, and P. Niyogi for making available their algorithm code on the web. We also thank the reviewers for their insightful suggestions.

M.L.G. is a stockholder in R2 Technology/Hologic and received royalties from Hologic, GE Medical Systems, MEDIAN Technologies, Riverain Medical, Mitsubishi and Toshiba. It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest that would reasonably appear to be directly and significantly affected by the research activities.

VIII. References

[1] M.L. Giger, H. Chan, and J. Boone, “Anniversary Paper: History and Status of CAD and Quantitative Image Analysis: The Role of Medical Physics and AAPM,” Med. Phys. **35**, 5799-5820 (2008).

[2] J. Shiraishi, L.L. Pesce, C.E. Metz, and K. Doi, “Experimental Design and Data Analysis in Receiver Operating Characteristic Studies: Lessons Learned from Reports in Radiology from 1997 to 2006,” Radiology **253**, 822-830 (2009).

[3] H.L. Kundel, C.F. Nodine, E.F. Conant, and S.P. Weinstein, “Holistic Component of Image Perception in Mammogram Interpretation: Gaze-Tracking Study,” Radiology **242**, 396-402 (2007).

[4] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning (Adaptive Computation and Machine Learning)* (The MIT Press, 2006).

[5] A.R. Jamieson, M.L. Giger, K. Drukker, H. Li, Y. Yuan, and N. Bhooshan, “Exploring Nonlinear Feature Space Dimension Reduction and Data Representation in Breast CADx with Laplacian Eigenmaps and T-SNE,” Med. Phys. **37**, 339-351 (2010).

[6] M. Belkin, and P. Niyogi, “Laplacian Eigenmaps for Dimensionality Reduction and Data Representation,” Neural Comput. **15**, 1373-1396 (2003).

[7] L. van der Maaten, and G. Hinton, “Visualizing Data Using T-SNE,” J. Mach. Learn. Res. **9**, 2605, 2579 (2008).

[8] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," J. Mach. Learn. Res. **7**, 2399-2434 (2006).
[9] M.L. Giger, Z. Huo, M. Kupinski, and C.J. Vyborny, "Computer-Aided Diagnosis in Mammography" in *Handbook of Medical Imaging, Volume 2. Medical Image Processing and Analysis,* edited by M. Sonka and J.M. Fitzpatrick *(SPIE Press Monograph Vol. PM80)* (SPIE--The International Society for Optical Engineering, 2000).
[10] B. Sahiner, H. Chan, N. Petrick, R.F. Wagner, and L. Hadjiiski, "Feature Selection and Classifier Performance in Computer-Aided Diagnosis: The Effect of Finite Sample Size," Med. Phys. **27**, 1509-1522 (2000)·
[11] M.A. Kupinski, and M.L. Giger, "Feature Selection with Limited Datasets," Med. Phys. **26**, 2176-2182 (1999).
[12] W. Chen, R.M. Zur, and M.L. Giger, "Joint feature selection and classification using a Bayesian neural network with automatic relevance determination priors: Potential use in CAD of medical imaging" in *Medical Imaging 2007: Computer-Aided Diagnosis, 2007,* edited by M. Giger and N. Karssemeijer , Proc. SPIE 6514, 65141G–10. (2007)
[13] Ming Li, and Zhi-Hua Zhou, "Improve Computer-Aided Diagnosis With Machine Learning Techniques Using Undiagnosed Samples," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on **37**, 1088-1098 (2007).
[14] G.N. Lee, and H. Fujita, "K-means Clustering for Classifying Unlabelled MRI Data," in *Proceedings of the 9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications* (IEEE Computer Society, 2007), pp. 92-98.
[15] H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, 2009).
[16] P. Kuksa, P. Huang, and V. Pavlovic, "Efficient Use of Unlabeled Data for Protein Sequence Classification: A Comparative Study," BMC Bioinformatics **10 Suppl 4**, S2 (2009).
[17] G.E. Hinton, "To Recognize Shapes, First Learn to Generate Images," Prog. Brain Res **165**, 535-547 (2007).
[18] K. Drukker, M.L. Giger, C.J. Vyborny, and E.B. Mendelson, "Computerized Detection and Classification of Cancer on Breast Ultrasound," Acad. Radiol. **11**, 526-535 (2004).
[19] K. Drukker, K. Horsch, and M.L. Giger, "Multimodality Computerized Diagnosis of Breast Lesions Using Mammography and Sonography," Acad. Radiol. **12**, 970-979 (2005).
[20] K. Horsch, M.L. Giger, L.A. Venta, and C.J. Vyborny, "Computerized Diagnosis of Breast Lesions on Ultrasound," Med. Phys. **29**, 157-164 (2002).
[21] K. Drukker, N.P. Gruszauskas, and M.L. Giger, "Principal component analysis, classifier complexity, and robustness of sonographic breast lesion classification," in *Medical Imaging 2009: Computer-Aided Diagnosis, 2009,* edited by M. Giger and N. Karssemeijer . Proc. SPIE 7260, 72602B–72602B6 . (2009).
[22] Y. Bengio, O. Delalleau, N.L. Roux, J. Paiement, P. Vincent, and M. Ouimet, "Learning Eigenfunctions Links Spectral Embedding and Kernel PCA," Neural Comput. **16**, 2197-2219 (2004).
[23] L. van der Maaten, "Learning a Parametric Embedding by Preserving Local Structure" in *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics* (JMLR W&CP, 2009), pp. 384-391.
[24] H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," J. Educ. Psychol. **24**, 498-520 (1933).
[25] F.R.K. Chung, *Spectral Graph Theory* (American Mathematical Society, 1997).
[26] M. Belkin, and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," Advances in Neural Information Processing Systems 14 **14**, 585--591 (2001).
[27] I. Nabney, *Netlab* (Springer-Verlag, London, Berlin, Hedelberg, 2002).
[28] M. Kupinski, D. Edwards, M. Giger, and C. Metz, "Ideal Observer Approximation Using Bayesian Classification Neural Networks," IEEE Trans. Med. Imaging **20**, 886-899 (2001).
[29] T.H.B. Schölkopf, and A.J. Smola, "Kernel Methods in Machine Learning," Ann. Stat. **36**, 1171-1220 (2008).
[30] CE Metz, et. al. "Software Programs Available from the Kurt Rossmann Laboratories" (2010) on website: http://xray.bsd.uchicago.edu/krl/KRL_ROC/software_index.htm.
[31] B. Sahiner, H. Chan, and L. Hadjiiski, "Classifier Performance Prediction for Computer-Aided Diagnosis Using a Limited Dataset," Med. Phys. **35**, 1559 (2008).

[32] S. Holm, “A Simple Sequentially Rejective Multiple Test Procedure,” Scand. J. Stat. **6**, 65-70 (1979).

[33] Z. Sidak, “Rectangular Confidence Regions for the Means of Multivariate Normal Distributions,” J. Am. Stat. Assoc. **62**, 626-633 (1967).

[34] I. Guyon, U. von Luxburg, and R. Williamson, "Clustering: Science or Art? Towards Principled Approaches” in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, 2009).

[35] I. Foster, Yong Zhao, I. Raicu, and S. Lu, " Cloud Computing and Grid Computing 360-Degree Compared *" in Grid Computing Environments Workshop, 2008. GCE '08* (2008), pp. 1-10.